Personality Sensing: Detection of Personality Traits Using Physiological Responses to Image and Video Stimuli

RONNIE TAIB, Data61, CSIRO
SHLOMO BERKOVSKY, Centre for Health Informatics, Macquarie University
IRENA KOPRINSKA and EILEEN WANG, School of Computer Science, University of Sydney
YUCHENG ZENG, School of Psychology, University of Sydney
JINGJIE LI, University of Wisconsin–Madison

Personality detection is an important task in psychology, as different personality traits are linked to different behaviours and real-life outcomes. Traditionally it involves filling out lengthy questionnaires, which is time-consuming, and may also be unreliable if respondents do not fully understand the questions or are not willing to honestly answer them. In this article, we propose a framework for objective personality detection that leverages humans' physiological responses to external stimuli. We exemplify and evaluate the framework in a case study, where we expose subjects to affective image and video stimuli, and capture their physiological responses using non-invasive commercial-grade eye-tracking and skin conductivity sensors. These responses are then processed and used to build a machine learning classifier capable of accurately predicting a wide range of personality traits. We investigate and discuss the performance of various machine learning methods, the most and least accurately predicted traits, and also assess the importance of the different stimuli, features, and physiological signals. Our work demonstrates that personality traits can be accurately detected, suggesting the applicability of the proposed framework for robust personality detection and use by psychology practitioners and researchers, as well as designers of personalised interactive systems.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Applied computing →** *Psychology*;

Additional Key Words and Phrases: Personality detection, framework, eye tracking, GSR, field study

ACM Reference format:
Ronnie Taib, Shlomo Berkovsky, Irena Koprinska, Eileen Wang, Yucheng Zeng, and Jingjie Li. 2020. Personality Sensing: Detection of Personality Traits Using Physiological Responses to Image and Video Stimuli. *ACM Trans. Interact. Intell. Syst.* 10, 3, Article 18 (October 2020), 32 pages.
https://doi.org/10.1145/3357459

The reviewing of this article was managed by special issue associate editors Shimei Pan, Oliver Brdiczka, Andrea Kleinsmith, Yangqiu Song.

Authors' addresses: R. Taib, Data61 - CSIRO, 13 Garden St, Eveleigh NSW 2015, Australia; email: ronnie.taib@data61.csiro.au; S. Berkovsky, Australian Institute of Health Innovation, Macquarie University, 75 Talavera Rd, North Ryde, NSW 2113, Australia; email: shlomo.berkovsky@mq.edu.au; I. Koprinska and E. Wang, School of Computer Science, University of Sydney, NSW 2006, Australia; emails: irena.koprinska@sydney.edu.au, ewan9058@uni.sydney.edu.au; Y. Zeng, School of Psychology, University of Sydney, NSW 2006, Australia; email: yzen7770@uni.sydney.edu.au; J. Li, Wisconsin Embedded Systems and Computing Lab, University of Wisconsin.Madison, 4613 Engineering Dr, Madison, WI 536706, USA; email: jingjie.li@wisc.edu.

2160-6455/2020/10-ART18 $15.00
https://doi.org/10.1145/3357459

1 INTRODUCTION

Personality is an established research area in psychology. A broad definition of personality refers to a set of individual patterns of behaviours, cognitions, and emotions that predict a human's behaviour and their interaction with the environment [66]. Although multiple personality theories have been developed, most of them conceptualise these patterns into *traits*, which are believed to be stable and consistent predictors of behaviour [50]. However, no single and agreed-upon model exists within the trait-based representation of personality, so multiple models, e.g., the Five Factor Model (known as the Big-5) [51], HEXACO [4], Temperament and Character Inventory [16], and Interpersonal Circumplex [17], have been developed and validated.

For any personality model considered, detection of the trait values is a complex and error-prone task. This is traditionally carried out in psychology research using validated questionnaires (or inventories) aimed at uncovering the values of the traits [55]. However, the fixed and long nature of the inventories often restrict practical applications. For example, the Minnesota Multiphasic Personality Inventory (MMPI-2) consists of 567 binary questions and is expected to take approximately 1.5 hours [11]. Likewise, the 300-item Adjective Checklist contains 300 adjectives, out of which the respondents mark those that best characterise them [27]. Moreover, the inventories often contain numerous similar questions, which limits their combined reach.

Furthermore, due to privacy considerations [5, 24] or in high-stake situations like job recruitment [3, 90], people may not be willing to genuinely answer the inventories, providing rather the desired or false answers. Faking, response distortion, and self-deception phenomena, and how to overcome these, are hotly debated contemporary issues in psychology research [23, 91]. Although their reliability is far from perfect, inventories remain the standard and most widely used tool for personality detection tasks. However, the above limitations trigger an increasing interest in innovative methods for objective, reliable, and practical detection of personality traits [42, 49, 61], including methods that leverage physiological signals [1, 6, 18, 72, 78].

In this work, we tackle the challenge of detecting personality traits using humans' physiological responses to external stimuli. Indeed, the rich modality, high accuracy, and moderate costs of modern sensing technologies facilitate their deployment in a range of applications. We propose to use such technologies for capturing physiological responses—in this case, eye activity and skin conductivity—of the human body in response to external stimuli. As many of these physiological responses cannot be consciously controlled [15], we posit that they can be considered as reliable and genuine indicators of the human's reaction to the stimuli and to the emotions evoked by the stimuli, which we attribute to human's personality. Hence, in this work, we set out to study whether such physiological responses to stimuli can serve as predictors of personality traits.

To this end, we propose a generic framework for objective detection of personality traits. The main components of the framework include: (i) external stimuli that triggers physiological responses; (ii) sensing technology that captures the responses to the stimuli; (iii) data processing component that segments the responses and extracts the features required for personality detection; and (iv) machine learning component that predicts the values of the personality traits. We initially present the framework, discuss the roles of the above four components and possible ways to implement them, and outline the dependencies between the components.

Then, we proceed to a specific instantiation of the framework, in which we use affective image and video stimuli, and eye-tracking and skin conductance responses, to detect personality traits. We focus on three established personality models: the Dark Triad [59], the Reinforcement Sensitivity model (known as BIS/BAS) [14], and the HEXACO model [4] (extension of the famous Big-5 [51]). We elaborate on the methodology for the data collection and analysis and then present the obtained results. These demonstrate that the framework is capable of accurately predicting a broad range of personality traits, with the combinations of image and video stimuli, as well as of

the eye-tracking and skin conductance data, yielding higher predictive accuracy than comparable methods.

Hence, the contributions of this work are three-fold: First, we demonstrate how off-the-shelf sensing and machine learning methods can be combined into a *generic framework for detection of personality traits*. Second, we exemplify and evaluate a specific *instantiation of the framework using affective image and video stimuli and two physiological responses*. Third, our evaluation achieves notably *high predictive accuracy results*, some of which are associated with well-studied factors in human psychology. Research into personality detection using physiological responses, as embodied by our framework, has a promising future for building usable and reliable trait detection systems. Being based on objective measurements from readily available devices, our framework could be automated to allow content personalisation and the design of user-aware interactive intelligent systems. It also has the potential to simplify and streamline various human modelling tasks for researchers and practitioners.

2 RELATED WORK

Personality is an organised set of characteristics that influences the individual's behaviours, cognition, and emotions [13, 66]. Modern personality theories conceptualise these characteristics into traits, which are believed to be relatively stable and consistent dispositions that humans possess. Within the trait-based representation of personality, multiple personality models have been proposed and studied. Influential and numerously validated models include the Big-5 Factor model [51] and its extension referred to as the HEXACO model [4], the Reinforcement Sensitivity Model (BIS/BAS) [14], the Dark Triad (D3) [59], and Interpersonal Circumplex [17].

One of the most widely accepted personality detection methods entails the administration of personality inventories [55]. These are questionnaires developed and validated based on relevant personality and psychometric theory. A self-reported questionnaire is typically used to measure personality traits and their facets. Although being easy to administer and process, the self-reported results can be distortion-prone, especially in high-stake situations [3, 23, 90]. This triggers an increasing body of research seeking for alternative distortion-resistant methods of personality detection. The popularity of social media opens the opportunity to detect personality through analysis of network activities and content posted by users [75, 76]. For example, Big-5 traits were linked to social network activity [26], while some traits were detected merely through the analysis of Facebook likes [42] and linguistic features of tweets [61]. Beyond social networks, deep learning was applied to detect the Big-5 traits from essays [49]. These methods, however, are not fully reliable, as people can often use social media for impression-management purposes [43, 70].

Personality traits may also impact the autonomic nervous system and, in turn, bodily responses and generated physiological signals [77]. To the best of our knowledge, the first significant attempt to detect the Big-5 traits using physiological signals was in References' [1, 18, 72, 78] lines of research. These works collected the electroencephalogram (EEG), galvanic skin response (GSR), face tracking, and electrocardiogram (ECG) data of subjects watching video clips. The obtained prediction accuracy levels varied substantially, ranging from below-random to 90%. However, this stream of research demonstrated the feasibility of personality detection using commercial-grade sensors. Beyond the Big-5 model, the 16 Personality Factors model was predicted using facial features and a neural network [25]. Although an improved accuracy was achieved, the method was complex and required about three hours for training the model.

Eye movement parameters were extensively used to detect conscious and unconscious activities. Complex features, such as gaze pattern and scan path, were found to be reliable indicators of cognitive strategies and attention [21, 62]. Pupillary response was used as an indicator of cognitive load [15, 84], whereas saccade amplitude and fixation durations were used for lie detection [48].

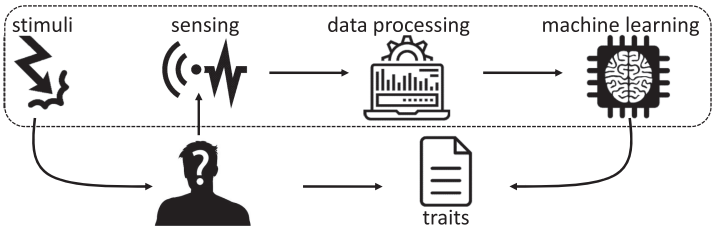

Fig. 1. Framework for personality detection.

In the personality domain, early research established links between eye contact, gaze aversion, and sociability [47]. With the advent of eye-tracking technologies, features derived from saccades, eye fixations, and pupils were found to be associated with personality traits [64, 82]. Recently, eye movement data during an everyday task was used to predict, although with relatively low levels of accuracy, the Big-5 traits and perceptual curiosity [35].

GSR measures the conductance of human skin, which can be seen as an indication of changes in sweat production, driven by the arousal of the sympathetic nervous system. Previous research linked GSR variation to stress, arousal, and cognitive load [40, 71]; as a human becomes more or less stressed, the skin conductance increases or decreases, respectively. Early work investigated the possibility of using GSR for recognising cognitive activity [54], while more recent worked also linked GSR to cognitive activity and stress [46] and established correlation between stress and cognitive load [15, 52]. Little work has focussed on GSR signal for personality detection, although links between the two were validated in References [1, 60] and recently in Reference [88].

This work addresses several limitations faced by prior research. Namely, we (i) extend the psychological traits to cover the D3, BIS/BAS, and HEXACO models; (ii) propose a framework for personality detection and exemplify it using eye-tracking and GSR data; (iii) examine two types of affective stimuli—images and videos—carefully crafted using valence-arousal and emotion metrics; (iv) experiment with a range of machine learning methods to optimise predictive accuracy; and (v) associate the obtained results to prior findings in psychology research.

3 PERSONALITY DETECTION FRAMEWORK

We start by presenting our framework for detection of personality traits using physiological signals. The framework is schematically depicted in Figure 1, where the components of the framework are placed within the dotted box. In the following paragraphs, we elaborate on these components.

3.1 External Stimuli to Trigger Physiological Responses

The main idea underpinning the proposed framework is that not consciously controllable physiological bodily responses to external stimuli can be considered as objective indicators of the subject's personality traits. Hence, the stimuli play an important role in triggering the desired responses.

A broad range of stimuli can be applied: from reading plain text, through watching multimedia documents, to carrying out interactive tasks. In this context, we highlight the desired links between the nature of the applied stimuli and the personality traits being detected. For example, when detecting a subject's learning style, a suitable stimulus could be a series of puzzles to solve or a set of cognitive tasks having an increasing degree of difficulty. On the contrary, detection of emotional stability may involve playing highly energetic songs or showing affective images with a strong graphic content. Hence, we posit that the selection of the appropriate stimuli needs to be tailored to the target psychological model and the target traits.

It is also important to consider the modalities of the stimuli and their presentation interface, as these are likely to cause different physiological responses. For example, a video recording of a car accident is expected to trigger stronger physiological responses than the textual description of the same accident, regardless of the phrasing of the text. Moreover, virtual reality presentation of exactly the same accident can further amplify the responses. To this end, when designing the stimuli, it is important to (i) trial and calibrate the impact of different modalities and stimuli; (ii) consider the order of the stimuli and their potential carry-over effects; and (iii) introduce cool-down periods between the presentation of the stimuli that can diminish such effects.

3.2 Sensing Technologies to Capture Physiological Responses

To capture physiological responses, the framework directly interacts with the subject as they are exposed to the stimuli. The responses are captured by the sensing technology in place and fed back into the framework for subsequent data processing and trait predictions.

Many options exist for the selection of the sensing technology used to capture physiological responses. However, revisiting the main idea of objective personality detection, we highlight the importance capturing responses that are not consciously controllable by the subject. Indeed, some responses, e.g., breathing rate or mouse movement patterns, may be controlled, while others, e.g., blinking rate or heart rate, may be controlled indirectly or only to a limited extent. However, numerous physiological responses, e.g., skin conductance, brain activity, or pupil size, cannot be controlled at all [15]. The level of control over the captured responses a subject can exhibit drives the objectivity level of the captured responses and, in turn, the reliability of the trait detection. We believe that the importance of such objectivity is paramount for reliable personality detection and highlight this as one of the main considerations affecting the selection of the sensing technology.

Another practical consideration refers to the usability and deployment of the sensors. Needless to say, the technology should accurately and reliably capture the target physiological responses and be able to mitigate the effect of external noises, e.g., those caused by temperature, subject's movements, and so forth. It is also desirable for the selected technology to be as unobtrusive and compact as possible, not to interfere with the normal interaction of the subject with the stimuli. Last, practical deployment aspects should also be considered. For example, the high costs and complex operation of fMRI sensors may restrict their practical *in situ* application for capturing brain responses, no matter how accurate they are.

3.3 Response Data Processing and Feature Extraction

The captured physiological responses will, for many sensing technologies, encapsulate raw signals, e.g., skin conductance values, electric signal produced by the brain, or heart rate records. To carry out meaningful predictions of personality traits, these signals need to be processed and features characterising the signals need to be extracted.

This component of the framework applies statistical and signal processing methods to process the captured physiological responses and extract predictive features, to be used by the subsequent machine learning component. The exact data processing steps largely depend on the selected sensing technology and captured responses. However, we identify three typical data processing steps: filtering, segmentation, and normalisation [69]. First, filters are applied to the captured data to mitigate various types of noise, e.g., those caused by the subject's body motion or degraded sensing quality. Then, temporal segmentation is applied to the filtered signal to split the data according to the desired time windows, e.g., periods when various stimuli were applied or when the subject rested between the stimuli. Finally, the segmented data are normalised with respect to an established baseline measure or the segment calibration data to diminish variance across individual recordings.

Data processing is followed by feature extraction. The features mostly depend on the selected technology. For most physiological signals, e.g., EEG and GSR data, the extracted features can be grouped into temporal and transform features. Temporal features represent the variability of the signal in the amplitude and frequency domains and include features such as the minimum, maximum, and mean signal energy, changes in frequency, and variance of the amplitude. Fourier or wavelet transform is typically applied to the captured signal to decompose it into the low-frequency and high-frequency components. Then, transform features characterising the signal in each component can be extracted.

3.4 Machine Learning for Personality Detection

The last component of the framework deals with the detection (or prediction) of personality traits. This is done using standard machine learning paradigms, where the learning component is trained on historical labelled data, e.g., from a pool of past subjects, and predicts labels for unknown data, i.e., traits of a new target subject, whose traits are being detected.

A broad range of supervised machine learning methods are applicable for this task. Generally, these are trained on a set of data labelled with the correct personality trait—the extracted features of subjects with known personality traits, where the trait values of the training data subjects serve as the labels for the feature vectors. Then, the values of the same traits are predicted for new target subjects, whose personality is unknown, given only the feature vectors of these subjects. Among popular supervised machine learning methods that can be applied for this task are decision trees (the trait value is predicted after inspecting the values of the extracted features individually and constructing if-then rules), regression models (the trait value is represented as a weighted linear combination of the values of the extracted features), and ensembles of classifiers that combine the predictions of several individual models.

It is also important to mention two potential challenges that may impede the application of machine learning methods. The first one is the lack of sufficiently large training data. Due to the sensitive and often complex nature of capturing physiological responses of humans, the data collection will mostly occur in a lab environment. This limits the volumes of data that can practically be collected and, as such, machine learning methods that can operate on limited training data may be preferred. The second challenge refers to the multitude of extracted features. For example, an EEG sensor may capture brain signals on 14 channels. If these are segmented and a number of features is extracted for every segment-channel combination, the overall number of features may quickly scale up to the thousands, which will potentially lead to overfitting. In this case, appropriate feature selection may need to be applied to select a smaller set of informative input features [86].

4 SETTING AND METHODS

This work describes an instantiation of the above framework. Namely, we detect personality traits using the eye-tracking (ETG) and *GSR data*, reflecting autonomic nervous activity elicited by physiological responses to *affective image and video stimuli*. The following subsections outline the personality models and traits, discuss the details of the framework application (video and image stimuli, sensors, feature extraction and selection, and trait classifiers), and finally outline the data collection methodology and evaluation setting.

4.1 Personality Models and Traits

We focus on three well-validated models: D3, BIS/BAS, and HEXACO. Table 1 briefly presents these models, their traits, and the relevant facets included in each model. Altogether, we examined 16 variables capturing the traits. To establish ground truth values for the 16 variables used as the class labels for the machine learning component, we deployed five well-validated personality

Table 1. Summary of the Studied Personality Models and Traits

Trait	Description
Dark Triad (D3)	
Primary Psychopathy	Primary psychopathy is the emotional aspect of psychopathy, characterised by a lack of empathy and deficit in processing negative feelings. It is associated with callousness, remorseless, and failure to accept responsibility [20].
Secondary Psychopathy	Secondary psychopathy is the behavioural aspect of psychopathy, characterised by antisocial acts. It is associated with instability and aggression, although it does not arise from deficit in processing negative feelings [20].
Tactics	Tactics is a component of Machiavellianism that focuses on exploitation of others. People high in tactics tend to engage in interpersonal exploitation, willingly and skilfully manipulating their peers in pursuit of personal goals [65].
Views	Views is a component of Machiavellianism that focuses on the lack of trust. People high in views hold a cynical view of the human nature, have a hyper-vigilance to being manipulated, with a view that others cannot be trusted [65].
Morality	Morality is a component of Machiavellianism that focuses on disbelief in the moral norms. People high in morality (more precisely, immorality) disregard conventional morality of the society, which would condemn their actions [65].
Narcissism	Narcissism involves excessive self-love. People high in narcissism have inflated sense of self-importance and self-admiration, with tendencies toward grandiose ideas, fantasied talents, and defensiveness to criticism [56].
Behavioural Inhibition System and Behavioural Activation System (BIS/BAS)	
BIS	BIS measures the motivation to avoid aversive outcomes. BIS is responsible for the experience of negative feelings such as fear, frustration, and sadness in anticipation for punishment. People high in BIS are more prone to anxiety [14].
BAS Drive	BAS Drive measures the motivation to persistently pursue the desired goals. People high in BAS Drive are more eager to engage in goal-directed efforts and to pursue their goals with perseverance [14].
BAS Fun Seeking	BAS Fun Seeking measures the motivation to find novel rewards spontaneously. People high in BAS Fun Seeking have a stronger desire for new rewards and a willingness to approach rewarding events on the spur of the moment [14].
BAS Reward Responsiveness	BAS Reward Responsiveness measures the sensitivity to pleasant reinforcers in the environment. People high in BAS Responsiveness are sensitive to rewards and positive stimuli and positively respond to the anticipation of reward [14].
HEXACO Personality Traits	
Agreeableness	Agreeableness concerns how people interact and maintain relationships with others. People high in agreeableness build warm relationships, are empathetic, altruistic, good-tempered, and less prone to conflicts [67].
Conscientiousness	Conscientiousness relates to the people's will to achieve their goals. People high in conscientiousness are more diligent, dutiful, organised, self-disciplined, and strive for achievements [67].
Extraversion	Extraversion relates to the sociability and assertiveness of people. People high in extraversion are sociable, gregarious, and seek excitement in interpersonal interactions with others [67].
Honesty	Honesty is associated with humility and sincerity of people. People high in honesty are generally loyal, truthful and direct, less hypocritical, less manipulative, and less deceitful [67].
Resiliency	Resiliency (or, the inverse trait, Neuroticism) concerns the emotional stability of people. People high in resiliency are better at emotional control, less impulsive, and less prone to anxiety and depression [67].
Openness	Openness is associated with people's acceptance of experiences and their creativity. People high in openness are more creative, curious, and have a stronger desire for novel experiences and intellectual exploration [67].

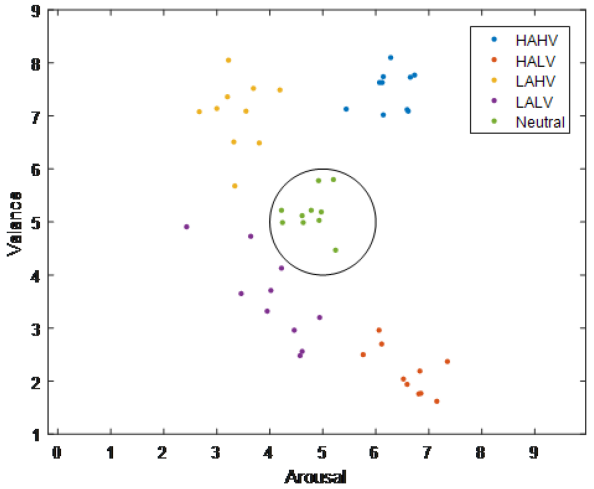

Fig. 2. The arousal-valence scores of the selected images.

inventories. For the six D3 traits, we use three inventories: (i) Levenson's Self-report Psychopathy inventory, which includes 26 items assessing psychopathy, both primary and secondary [45]; (ii) Narcissistic Personality Inventory (NPI-16), which is a short 16-item version of the full NPI-40 inventory measuring narcissism [2]; and (iii) MACH-IV, which is a trimmed 20-item inventory extracted from the full MACH inventory measuring Machiavellianism, including the tactics, morality, and views traits [63]. The four BIS/BAS traits (BIS, BAS Drive, BAS Fun Seeking, and BAS Reward Responsiveness) are measured using the BIS/BAS inventory containing 24 items [28], whereas the six HEXACO traits (agreeableness, conscientiousness, extraversion, honesty, resiliency, and openness) are measured using a 25-item inventory [74]. It should be highlighted that all the inventories used in this work are well-studied and validated tools, used in personality research for a range of personality detection tasks.

A score for the above 16 variables for each subject is discretised and used as a class label for the recorded ETG and GSR signals to create training data for the trait classifiers. The classifiers are trained to predict the trait values from the collected data and are then used to determine the trait class label for a new subject. We will detail the performance evaluation metrics later in this section.

4.2 Affective Image and Video Stimuli

In this work, we applied affective stimuli, expected to evoke the subjects' emotional responses. In particular, we opted for still images and short video stimuli, because we wanted to develop a system able to provide a fast psychological profiling.

Images. We used a subset of images from the International Affective Picture System (IAPS) dataset [44]. This is a well-studied dataset, where each image is associated with numeric scores corresponding to different emotions. We focused on the arousal and valence scores and clustered images into five groups: high arousal and high valence (HAHV, strongly positive emotions), low arousal and high valence (LAHV, mildly positive emotions), low arousal and low valence (LALV, mildly negative emotions), high arousal and low valence (HALV, strongly negative emotions), and neutral images (neutral emotions, no specific arousal applies). A set of 50 IAPS images—10 for each group—was selected as the image stimuli. Figure 2 shows the arousal-valence scores of the selected

images.[1] As can be seen, the selected images demonstrate low intra-cluster distance, in particular the two groups of high-arousal images.

The images from each group were shown for 8 seconds each, in blocks of five of the same arousal and valence, and in the same order of blocks (HAHV, LAHV, neutral, LALV, HALV) to all the subjects. Each block was preceded by a cool-down period of 15 seconds, during which a black cross on a white background was shown, allowing recovery from previous stimuli. The subjects could neither pause the presentation nor skip images, so the overall duration of the image stimuli was 9.5 minutes.

Videos. We used videos from the English version of the FilmStim dataset [68]. Video stimuli representing seven emotion types—fear, tenderness, anger, neutral, sadness, amusement, and disgust—were selected based on their pre-annotated arousal-valence scores. These video stimuli were extracted from the following movies: "Seven," "Life is Beautiful," "American History X," "Blue," "Dangerous Mind" "A Fish Called Wanda," and "Trainspotting." The duration of the videos was between 25 and 132 seconds to ensure that the emotional peak is reached while avoiding subject fatigue.

The videos were shown in the same order to all the subjects and broken down into two blocks to minimise carry-over effects. The first block included videos evoking fear, tenderness, and anger, while the second included videos evoking neutral emotions, sadness, amusement, and disgust. A cool-down period of 30 seconds was allocated after each video, during which a black cross on a white background was displayed. The overall duration of the video stimuli was 14 minutes.

4.3 ETG and GSR Sensors

We used SMI eye-tracking glasses[2]—lightweight wearable glasses able to capture natural eye and gaze behaviour through two infrared cameras focusing on each eye. A relatively wide field-of-view angle is captured: 60° horizontally and 46° vertically. Eye data are estimated in real-time and transmitted to a server storing the data and producing various metrics. Eventually, the sensor captures and provides commonly used eye data, such as pupil dilation (along X and Y axes), eye saccades and fixations, blinks, and relative gaze direction. The accuracy and availability of these features depend on the duly controlled illumination conditions and subject movement.

We collected GSR data using a Procomp Infiniti[3] integrated biometrics acquisition device. This measures skin conductance created by micro-sweating, controlled by the sympathetic nervous system [71, 87]. It is connected to the experiment computer via an optic fibre link and USB adapter to prevent electromagnetic noise, and the raw data is timestamped using the computer clock with a millisecond resolution. The GSR data are post-processed and synchronised with ETG and other task data using synchronisation blocks at the start and end of the experiment.

Figure 3 shows the actual experiment setup with a subject wearing the ETG and GSR sensors. We also captured EEG responses, but in this analysis below, we focus only on ETG and GSR as sensing technologies that can be deployed unobtrusively in practical settings.

4.4 Data and Features

Features extracted from the captured ETG data relate to eye activity and can be categorised into three measurement groups: eye blink, eye movement (saccades and fixations), and pupillary response. Specifically, we extracted ten ETG features listed in the top part of Table 2. As for the GSR signal, we expanded traditional GSR measurements with features reflecting the power and statis-

[1] Due to the terms of use of IAPS, samples of the images are not included.
[2] https://www.smivision.com/.
[3] http://thoughttechnology.com/index.php/procomp-infiniti-320.html.

Fig. 3. Lab data collection setup.

tical characteristics of the raw signal, so we extracted nine GSR features listed in the bottom part of Table 2. These features are in line with the features used in prior works [15, 57, 85].

The above features were populated for 17 temporal blocks. The images were shown in ten blocks, as two sequences of five blocks corresponding to the arousal-valence groups: HAHV, LAHV, LALV, HALV, and neutral. In addition, each video (fear, tenderness, anger, neutral, sadness, amusement, and disgust) was considered as an individual temporal segment. Hence, the segmentation splits the signal into ten temporal image blocks and seven video blocks. Cool-down periods between the blocks were used as the baseline data.

For the ETG signal, we focussed on blinks, fixations, saccades, and pupil features, which exhibit user-specific characteristics and vary over the time of day. Hence, the process started by a min-max normalisation of the raw feature values in each temporal block with respect to the baseline observed in the cool-down period immediately preceding the block. This was done to ensure each block can be compared to other blocks in different conditions. Then, all features were segmented according to each block.

For GSR, high-frequency noise is traditionally removed using a low-pass filter; hence, a 5-order Butterworth filter was used to filter out the frequency components above 0.5 Hz. Then, the signal mean and Hjorth parameters were calculated for each block [34] to extract time-based features. Hjorth parameters are popular in EEG signal processing, as they extract a small set of slope-related features from non-stationary signals with irregular or non-uniform shapes. The energy of the signal was computed using the Welch's power spectrum density estimation [80]. We applied convex optimisation to break the signal into the tonic and phasic components [29, 30] and we also considered the area under the decomposed phasic signal.

4.5 Feature Selection

As the stimuli were split into 17 temporal blocks, and ten ETG and nine GSR features were extracted for each block, the overall number of features was in the hundreds. The sheer number of features also introduces the risk of overfitting, i.e., situation where the training of the trait predictors overuses subset of features that are good for the specific training data and may not scale beyond this [41].

Table 2. Extracted ETG and GSR Features

ETG features	
Blink Rate (BR)	Average number of blinks per second; blink count divided by block duration.
Saccade Rate (SR)	Average number of saccades per second; saccade count divided by block duration.
Saccade Amplitude (SA)	Average angular distance of the saccades (in °), over all saccades in the block.
Average Saccade Velocity (ASV)	Average angular velocity of the saccades (° per second), over all saccades in the block.
Peak Saccade Velocity (PSV)	Average of the peak angular velocities (° per second) of each saccade in the block.
Fixation Rate (FR)	Average number of fixations; fixation count divided by block duration.
Fixation Duration (FD)	Average duration of fixations; cumulative duration of fixations divided by their count.
Saccade-Fixation Ratio (SFR)	Ratio between the duration of saccades (search) and fixations (processing) in the block.
Horizontal pupil size (PX)	Average horizontal diameter (in pixels) of the pupils over the duration of the block.
Vertical pupil size (PY)	Average vertical diameter (in pixels) of the pupils over the duration of the block.
GSR features	
Signal mean (SM)	Mean value of the raw skin conductance signal over the duration of the block.
Signal energy (SE)	Mean energy of the raw signal estimated using Welch's power spectral density.
Nerve activations rate (NAR)	Average number of sudomotor nerve activations over the duration of the block.
Hjorth activity (HA)	The activity parameter, representing the signal power, variance of a time function.
Hjorth mobility (HM)	The mobility parameter, representing the mean frequency of the signal.
Hjorth complexity (HC)	The complexity parameter, representing temporal changes in frequency.
Phasic peaks rate (PPR)	Average number of phasic peaks over the duration of the block.
Phasic mean amplitude (PMA)	Mean value of the phasic peak amplitude over the duration of the block.
Phasic area (PAR)	Value of the integrated phasic component over the duration of the block.

To mitigate the risk of overfitting, we conducted feature selection using the Correlation-based Feature Selection (CFS) algorithm [32]. The main idea underpinning CFS is that a good feature subset should contain features that are highly correlated with the class label, i.e., very informative for the predictions, but weakly correlated with other features, i.e., not redundant. Essentially, CFS defines a heuristic measure based on these two criteria and uses a search algorithm to find the feature subset that maximises this measure.

CFS was applied for predictions of each trait individually and was found to substantially reduce the number of features used for the predictions. Table 3 summarises the number of selected features for the ETG and GSR signals. We observe that the number of selected features was between three and ten for ETG, and between five and nine for GSR. This accounts for a considerable feature set reduction of over 91% for the ETG signal and over 94% for GSR, and reduces the risk of overfitting.

4.6 Trait Prediction Models

As our subjects were not recruited on psychological grounds, we assume that they were unlikely to exhibit extremely low or high trait values and that their trait values distribute normally. Hence,

Table 3. Sample Feature Selection Results

Signal	Selected	Smallest set	Largest set	Reduction
ETG	3–10	BAS Reward Responsiveness	Openness	91.5%–97.5%
GSR	5–9	Secondary Psychopathy, BIS, Agreeableness, Honesty, Openness	BAS Reward Responsiveness, Conscientiousness, Extraversion	94.1%–96.7%

Table 4. Parameterisation of Weka Classifiers

Classifier	Weka Name	Parameters
AB	meta.AdaBoostM1	tree type: DT (J48), num. of iterations (trees combined): 10
DT	trees.J48	confidence factor for pruning: 0.25, min. num. of instances per leaf: 2
LR	functions.Logistic	ridge value: 1.0E-8
NB	bayes.NaiveBayes	normal distribution probability density for numeric features
RF	trees.RandomForest	num. of iterations (trees combined): 100, num. of randomly selected features: int(log_2(num. of predictors)+1)
SVM	functions.SMO	kernel: polynomial, complexity parameter: 1, epsilon: 1.0E-12, tolerance: 0.001
kNN	lazy.IBk	num. of neighbours: 3, distance: Euclidean

raw trait values obtained through the personality inventories were discretised into three classes—low, medium, and high—for each trait, using equal-frequency binning. This introduced potential risk of discretising subjects with similar trait values into different classes. However, this allowed us to represent the trait prediction task as a classification problem and use the discrete class labels for training and predictions.

To deploy the predictors, we used standard implementations of seven classifiers offered by Weka, an open-source data mining toolbox [33]. Specifically, the following seven classifiers were deployed: AdaBoost (AB), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbour (kNN). As we used standard off-the-shelf implementations of established classification methods, we do not include the descriptions of these algorithms and refer the reader to Reference [33] for further details. We used the default Weka's parameterisation for all the classifiers, except for (i) the number of neighbours in kNN was set to k = 3, and (ii) decision trees rather than decisions stumps were combined in AB. The key parameters of the deployed classifiers are detailed in Table 4.

A separate classifier was trained for the predictions of each trait, such that we ended up with 16 classifiers. One training data point corresponds to one subject and includes the values of the selected features and the trait label assigned based on the discretised scores for the trait. Due to the equal-frequency binning, the classification problem was class-balanced, as the number of training data points in every class, and for every trait, was identical. Given the block features of the training subjects, their trait labels, and the block features of the target subject, the goal of the classifier was to predict the trait class label for the target subject.

To evaluate the performance of the classifiers, we applied the leave-one-out methodology, the most appropriate methodology for our data, which allows to obtain accurate performance estimates even for small and sparse datasets [83]. This involves repeated runs using the data of all but

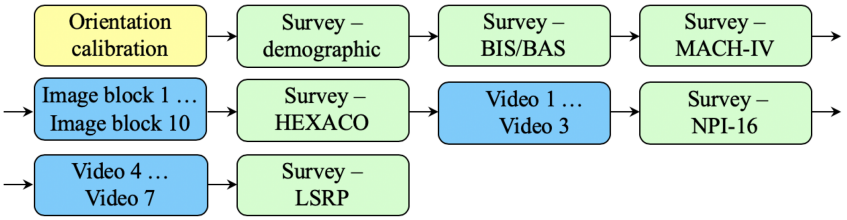

Fig. 4. Experimental workflow: stimuli and inventories.

one subject to train the classifiers, while the data of the last target subject are used for testing. At each run, the goal of the classifier was to predict the trait class label for the target subject. The accuracy of the prediction was assessed against the withheld class label of the trait for this subject. The process was repeated using another subject as a target until all subjects were tested. Finally, the average accuracy over all 21 runs was calculated and reported.

4.7 Data Collection and Evaluation Setting

In total, 21 subjects were recruited for this experiment, under ethics clearance from a nationally accredited committee. The subjects were either students or staff of a research organisation. All but one subject reported good or native English proficiency. Eighteen subjects were aged 18 to 30, whereas the other 3 were older than 30. The data were collected in a controlled laboratory setting under fixed illumination and room temperature conditions (see Figure 3), individually for every subject.

The workflow of the conducted data collection is shown in Figure 4. The whole procedure took, on average, 55 minutes. At the start, the subjects were provided with an overview of the experiment and their written consent was obtained. The eye-tracking glasses and skin conductance sensor were then mounted and calibrated to achieve the best signal quality. The subjects were instructed to sit down and relax, to acquire a baseline signal, and minimise artefacts related to movement.

The ground truth data, also used as the target class label for the traits, was obtained by administering the personality inventories for the 16 traits listed in Table 1. Note that the inventories were interleaved with the image and video stimuli to avoid fatigue and provide additional cool-down time that would further diminish carry-over effects. As shown in Figure 4, there were six inventories (marked in green). The average cumulative completion time of these inventories was 11 minutes.

We show in Table 5 the descriptive statistics of the raw scores obtained for the 16 personality traits. The data include the overall range of values in the inventory, the mean and standard deviation obtained for the 21 subjects, and the brackets of the low, medium, and high classes within each trait. Despite the relatively small sample size, the observed ranges generally correspond to the established D3 and BIS/BAS norms, whereas there exists little evidence of the HEXACO norms [38, 45, 73]. Overall, very few scores are placed at the extremities of the trait ranges[4] although the equal-frequency binning ensures uniform distribution of subjects across the classes.

The subjects were exposed to the image and video stimuli, and the ETG and GSR sensors were used to capture the physiological responses. We processed the collected data, populated the features for the temporal blocks, and fed them into the classifiers, which were trained to predict the labels of the traits. We used the classification *accuracy* metric to evaluate the performance of the classifiers. This is the ratio between the number of correctly predicted trait class labels (low,

[4]It should be noted that this could have led to a biased training data and less reliable predictive accuracy for extreme values of the traits. We elaborate on this limitation of our work in the concluding sections of the article.

Table 5. Descriptive Statistics of Personality Scores (the Range Column Shows Theoretical Trait Ranges, while Low, Medium, and High Reflect Our Study Observations)

Trait	Range	Mean	SD	Low	Medium	High
Prim. Psychopathy	[16,64]	30.05	5.21	[21,26]	[28,32]	[33,40]
Second. Psychopathy	[10,40]	20.68	2.34	[17,19]	[20,21]	[22,25]
Tactics	[9,45]	23.91	3.96	[19,20]	[21,25]	[26,32]
Views	[9,45]	23.68	4.06	[17,21]	[22,24]	[25,32]
Morality	[2,10]	5.18	1.53	[2,4]	[5,5]	[6,8]
Narcissism	[0,100]	29.47	22.14	[0,13]	[18,32]	[37,82]
BIS	[7,28]	20.23	3.04	[15,18]	[19,20]	[21,26]
BAS Drive	[4,16]	11.00	1.67	[9,9]	[10,11]	[12,14]
BAS Fun Seeking	[4,16]	11.59	1.94	[8,10]	[11,12]	[13,15]
BAS Reward Resp.	[5,20]	15.64	1.89	[13,14]	[15,15]	[16,20]
Agreeableness	[4,20]	13.95	2.17	[10,13]	[14,14]	[15,18]
Conscientiousness	[4,20]	13.14	2.80	[9,11]	[12,14]	[15,20]
Extraversion	[4,20]	14.14	2.42	[11,12]	[13,14]	[15,19]
Honesty	[5,25]	17.91	2.91	[14,15]	[16,19]	[20,24]
Resiliency	[4,20]	13.64	2.63	[9,12]	[13,15]	[16,19]
Openness	[4,20]	13.82	2.86	[9,12]	[13,14]	[16,19]

medium, or high) and the total number of predictions made by the classifier. We also computed the precision and recall of the predictions and combined them into the *F1 score*, which is the harmonic average of precision and recall. The reported statistical significance test results were produced by a Friedman test followed by post hoc pairwise Wilcoxon signed-rank tests, with a Bonferroni correction for the number of comparisons by using an alpha of $.05 \div 3 = .0167$ and $.05 \div 7 = .0071$ for comparison between three and seven groups, respectively. Such non-parametric tests are used due to the non-normal data distribution. Given the limited sample size, we also used the two-tailed exact test to best estimate the metrics.

5 RESULTS

The following questions guide our analysis:

—Q1: What machine learning methods can be used for detecting traits?
—Q2: What traits can be detected with high/low levels of accuracy?
—Q3: What type of stimuli is most informative for trait detection?
—Q4: What features are most predictive of each personality model?
—Q5: What physiological signal is most informative for trait detection?

5.1 Classifier Performance

Q1 deals with the performance of various machine learning methods applied for trait detection. We assess every classifier through the mean accuracy and mean F1 scores, both computed across all the traits and all the subjects. Table 6 shows the performance of the seven classifiers using both the image and video stimuli. The best-performing method is highlighted in bold.

Considering the ETG signal, we observe that NB, the most accurate classifier, achieves acc = 0.860, and outperforms other classifiers by 9.1% or more. LR, the second-best classifier, achieves acc = 0.789, followed by SVM and kNN, achieving acc = 0.756 and acc = 0.726, respectively. The remaining classifiers are substantially lower—achieving accuracy between 0.351 and 0.631. F1 scores

Table 6. Performance of the Seven Classifiers

Signal	Metric	AB	DT	LR	NB	RF	SVM	kNN
ETG	Accuracy	0.539	0.351	0.789	**0.860**	0.631	0.756	0.726
	F1 score	0.535	0.344	0.784	**0.860**	0.625	0.752	0.714
GSR	Accuracy	0.542	0.241	0.792	**0.818**	0.625	0.762	0.732
	F1 score	0.539	0.234	0.787	**0.819**	0.618	0.758	0.727

exhibit the same trends as the accuracy: NB achieves the highest F1, which is higher than the second-best LR by 9.7%. Then come SVM and kNN, while the other classifiers are all substantially less accurate.

Turning to the GSR signal, we observe similar results, where NB achieves slightly lower accuracy and F1 scores. NB still outperforms LR, the second-best classifier by 3.4% in terms of accuracy and by 3.9% in terms of F1. Like for the ETG signal, the third-best classifier is SVM, which is marginally inferior to LR both for accuracy and F1. These three are followed by kNN, while the remaining classifiers all achieve accuracy and F1 scores below 0.625.

The differences in ETG accuracy and F1 score are statistically significant (Friedman $p < .001$ for both). The pairwise post hoc comparisons of NB against each other classifier show that the superiority of NB is statistically significant for both accuracy and F1 score (Wilcoxon $p < .0071$, Bonferroni correction for seven tests). Similarly, the differences in GSR accuracy and F1 score are statistically significant (Friedman $p < .001$ for both). The pairwise post hoc comparisons for GSR also show that the superiority of NB is statistically significant (Wilcoxon $p < .0071$), except against LR ($p = .141$ for accuracy and $p = .078$ for the F1 score) and SVM ($p = .027$ for accuracy and $p = .056$ for F1).

The results using the ETG and GSR signals clearly demonstrate that the NB classifier performs well in combination with CFS, a correlation-based feature selection method. NB's performance is adversely affected by highly correlated input features, since it assumes that the features are independent from each other within the class [32]. By applying CFS prior to the classification, we select a subset of less correlated features, which aligns with NB's underlying assumption and contributes to its predictive performance. Likewise, LR achieves high classification accuracy, when preceded by the CFS feature selection of non-correlated input features [36, 83]. The DT classifier achieves the lowest performance. A closer examination shows that the generated trees are shallow, examining 4–5 features only, selected based on their information gain, which is insufficient for the target classification task. As expected, tree ensembles implemented by AB and RF improve the performance of DT. However, their performance is still not competitive with NB and LR. The similarity- and separation-based kNN and SVM classifiers achieve better performance than the tree-based methods but are still inferior to NB.

Hence, revisiting Q1, we conclude that for our dataset and personality detection task, *Naive Bayes is the most appropriate machine learning method* out of the seven classifiers under investigation. In the following experiments, we will primarily focus on the results achieved by the NB classifier.

5.2 Individual Trait Detection

Q2 deals with the detection of individual personality traits. Table 7 shows the accuracy and F1 scores of the NB classifier using the ETG signal when predicting each of the 16 traits individually. For benchmarking purposes, we also show the highest accuracy achieved for the trait by any other classifier (this best-other classifier is named in brackets). The best-performing method for each trait is highlighted in bold. The average accuracy and F1 scores of NB for each psychological model and across all the 16 traits are also shown.

Table 7. Performance of NB and Best-other Classifiers Using the ETG Signal

	accuracy		F1 score	
Trait	**Best-Other**	**NB**	**Best-Other**	**NB**
Primary Psychopathy	**0.762** (kNN)	**0.762**	**0.761** (kNN)	0.756
Secondary Psychopathy	0.810 (SVM)	**0.857**	0.812 (SVM)	**0.856**
Tactics	**0.905** (SVM)	**0.905**	**0.905** (SVM)	**0.905**
Views	**0.905** (LR,SVM)	**0.905**	**0.904** (SVM)	**0.904**
Morality	0.810 (LR)	**0.905**	0.798 (LR)	**0.906**
Narcissism	0.762 (LR)	**0.810**	0.760 (LR)	**0.812**
Mean D3		*0.857*		*0.856*
BIS	**0.905** (LR)	**0.905**	**0.904** (LR)	**0.904**
BAS Drive	**0.857** (SVM)	0.810	**0.853** (SVM)	0.819
BAS Fun Seeking	**0.952** (SVM)	**0.952**	**0.952** (SVM)	**0.952**
BAS Reward Responsiveness	**0.905** (SVM)	0.857	**0.897** (SVM)	0.850
Mean BIS/BAS		*0.881*		*0.882*
Agreeableness	0.857 (LR)	**0.905**	0.856 (LR)	**0.905**
Conscientiousness	0.714 (LR)	**0.810**	0.717 (LR)	**0.810**
Extraversion	0.714 (LR)	**0.810**	0.716 (LR)	**0.811**
Honesty	0.762 (LR)	**0.810**	0.762 (LR)	**0.810**
Resiliency	0.762 (AB,LR,RF,kNN)	**0.905**	0.760 (AB,RF)	**0.905**
Openness	**0.857** (LR)	**0.857**	0.856 (LR)	**0.858**
Mean HEXACO		*0.849*		*0.850*
Mean Overall		*0.860*		*0.860*

Comparing NB with the other best-performing classifier, we observe several trends. NB yields the highest accuracy for eight traits, for two traits SVM outperforms NB, and in six cases NB is as good as the other best-performing classifier. The same observations are strengthened considering the F1 scores, as NB outperforms the other classifiers for nine traits. We also analyse the performance of NB across the three psychological models. For D3, NB beats the accuracy of the best-other classifier for Secondary Psychopathy, Morality, and Narcissism, while LR and SVM are the dominant best-other classifiers. For BIS/BAS, SVM outperforms NB for BAS Drive and BAS Reward Responsiveness. For HEXACO, NB clearly dominates other classifiers, beating them for all the traits but Openness. Here, LR is the best-other classifier, coming second for five traits. Overall, LR is the best-other classifier for the ETG signal. The F1 scores largely mirror the accuracy observations. We do not test the statistical significance of these results, as NB is not compared against a specific algorithm, but rather against the best-performing other classifier, which varies across the traits.

Table 8 presents the accuracy and F1 scores of predictions of the 16 traits individually using the GSR signal. Considering the accuracy scores, NB is the best-performing classifier for 4 traits, performs on par with the best-other classifier for 8 more traits, and it is outperformed for 4 traits only. In terms of the F1 scores, NB is the best-performing classifier for 7 traits, while it is outperformed also for 7 traits, and for two the observed performance is similar. Here, we also do not report the significance scores, as for every trait NB is compared against a different best-performing algorithm.

Notably, for both the GSR and ETG signals, LR is slightly better than SVM in terms of accuracy. Inspecting the accuracy scores for the individual traits, LR is the best-other classifier for 10 traits for both signals, whereas SVM is the best for 6 traits for ETG and 7 traits for GSR. For the ETG

Table 8. Performance of NB and Best-other Classifiers Using the GSR Signal

	accuracy		F1 score	
Trait	**Best-Other**	**NB**	**Best-Other**	**NB**
Primary Psychopathy	**0.857** (SVM)	**0.857**	0.853 (SVM)	**0.857**
Secondary Psychopathy	**0.714** (LR)	**0.714**	0.717 (LR)	**0.727**
Tactics	**0.952** (LR,SVM)	**0.952**	**0.952** (LR,SVM)	**0.952**
Views	**0.857** (LR,SVM,kNN)	**0.857**	**0.859** (LR)	0.853
Morality	**0.810** (LR,kNN)	0.762	**0.810** (LR)	0.757
Narcissism	0.714 (LR,kNN)	**0.762**	0.713 (kNN)	**0.766**
Mean D3		*0.817*		*0.819*
BIS	**0.905** (LR)	**0.905**	**0.904** (LR)	**0.904**
BAS Drive	0.619 (kNN)	**0.667**	0.616 (kNN)	**0.655**
BAS Fun Seeking	**0.905** (SVM)	**0.905**	0.898 (SVM)	**0.904**
BAS Reward Responsiveness	**0.952** (SVM)	0.905	**0.953** (SVM)	0.903
Mean BIS/BAS		*0.845*		*0.842*
Agreeableness	**0.857** (LR)	0.762	**0.858** (LR)	0.769
Conscientiousness	**1.000** (SVM)	0.952	**1.000** (SVM)	0.952
Extraversion	0.667 (kNN)	**0.762**	0.682 (kNN)	**0.770**
Honesty	0.667 (LR,SVM)	0.667	**0.664** (SVM)	0.656
Resiliency	**0.857** (LR)	0.810	**0.849** (LR)	0.811
Openness	0.810 (LR)	**0.857**	0.810 (LR)	**0.857**
Mean HEXACO		*0.802*		*0.803*
Mean Overall		*0.818*		*0.819*

signal, only SVM outperforms NB, namely, for BAS Drive and BAS Reward Responsiveness. For the GSR signal, both LR and SVR aee able to outperform NB for some of the traits. The same trend holds for the F1 scores, where LR is the most represented second-best classifier, closely followed by SVM, respectively, for 8 and 6 of the ETG traits, and 7 and 6 of the GSR traits.

Combining the results in Tables 7 and 8, we re-affirm our earlier finding that NB is the most appropriate method for personality trait predictions. For all but 1 trait predicted with ETG, NB achieves acc = 0.810 or greater. For GSR it achieves an accuracy of at least 0.667 for all traits. This is twice better than the accuracy of a random guess in a three-class classification. Out of the 16 studied traits, NB predicts 9 traits with acc = 0.900 or greater, and additional 7 with accuracy between 0.800 and 0.900, considering both the ETG and GSR signals. Overall, the better performing ETG signal achieves mean acc = 0.857 for the D3 model, acc = 0.881 for BIS/BAS and acc = 0.849 for HEXACO. Considering the best-other classifier, we conclude (in line with Table 6) that LR is second-best for the ETG and GSR. Further refining this, we note that SVM can be a strong alternative to NB for BIS/BAS and potentially D3 traits, for both ETG and GSR signals.

To explain these results, we resort to the very nature of the predicted traits. Established psychology research classified personality traits into three broad categories: those driven by affect, by cognitions, or by behaviours. We group the latter two into the non-affective category. Specifically, References [79] and [37] associated Machiavellianism with the affective rather than cognitive assessment. As the deployed stimuli were validated affective images and videos, they presumably evoked emotional responses and, as a result, we observe the Tactics, Views, and Morality traits (the former with both ETG and GSR), all being predicted with acc > 0.900. Similarly, Reference [38] analysed the links between the BIS/BAS traits and affect and found BIS and BAS Fun Seeking

Table 9. Performance of the Image and Video Stimuli for the ETG Signal

	accuracy			F1 score		
Trait	**Image**	**Video**	**Both**	**Image**	**Video**	**Both**
Primary Psychopathy	**0.667**	0.619	**0.762**	**0.664**	0.624	**0.756**
Secondary Psychopathy	0.810	**0.857**	**0.857**	0.806	**0.860**	0.856
Tactics	**0.810**	0.667	**0.905**	**0.804**	0.649	**0.905**
Views	0.714	**0.905**	**0.905**	0.705	**0.904**	**0.904**
Morality	0.810	**0.857**	**0.905**	0.800	**0.857**	**0.906**
Narcissism	0.619	**0.762**	**0.810**	0.619	**0.748**	**0.812**
Mean D3	*0.738*	***0.778***	***0.857***	*0.733*	***0.774***	***0.856***
BIS	0.714	**0.762**	**0.905**	0.686	**0.755**	**0.904**
BAS Drive	0.667	**0.714**	**0.810**	0.654	**0.718**	**0.819**
BAS Fun Seeking	**0.857**	0.714	**0.952**	**0.857**	0.718	**0.952**
BAS Reward Responsiveness	0.571	**0.857**	0.857	0.558	**0.850**	0.850
Mean BIS/BAS	*0.702*	***0.762***	***0.881***	*0.689*	***0.760***	***0.882***
Agreeableness	**0.857**	0.762	**0.905**	**0.855**	0.766	**0.905**
Conscientiousness	0.619	**0.762**	**0.810**	0.603	**0.765**	**0.810**
Extraversion	**0.810**	0.619	**0.810**	**0.814**	0.614	0.811
Honesty	0.714	**0.810**	**0.810**	0.712	**0.806**	**0.810**
Resiliency	0.762	**0.810**	**0.905**	0.714	**0.804**	**0.905**
Openness	**0.762**	**0.762**	**0.857**	**0.763**	0.753	**0.858**
Mean HEXACO	***0.754***	***0.754***	***0.849***	*0.744*	***0.752***	***0.850***
Mean Overall	*0.735*	***0.765***	***0.860***	*0.726*	***0.762***	***0.860***

to be correlated with positive and negative affect, respectively. We observe that these two BIS/BAS traits are predicted with acc > 0.900 using both ETG and GSR, which, given the affective stimuli, is consistent with Reference [38]. Considering the HEXACO traits, Reference [92] identified Neuroticism, the Big-5's counter-part of Resiliency, to be the only trait associated with affect. Inspecting the results in Table 7, we find Resiliency and Agreeableness being predicted with acc > 0.900, while the predictions of other HEXACO traits are less accurate.

Hence, we summarise Q2 and conclude that the affective nature of the stimuli allows us to generate *more accurate predictions for traits associated with affect*, in several cases, regardless of the applied sensing technology. Other personality traits, associated with either behaviours or cognitions, are generally predicted with a lower degree of accuracy.

5.3 Image vs. Video Stimuli

Next, we turn to Q3 and assess whether image or video stimuli are more predictive of the personality traits. For this, we separate the signals captured in response to the image stimuli from those captured in response to the video stimuli and use them individually for feature extraction, classifier training, and trait predictions. Table 9 shows the mean performance obtained for each trait by the NB classifier and ETG signal, using either the image or video stimuli, and then performance of the combined image and video stimuli (the column "Both" is identical to the "NB" column in Table 7). The accuracy and F1 scores are shown separately, and the best-performing stimuli (or combination of stimuli) for every trait is highlighted in bold.

Focussing on the classification accuracy, we observe that the video stimuli achieved a higher overall mean accuracy than images, 0.765 vs. 0.735. This observation is valid for the mean scores

obtained for the D3 (0.778 vs. 0.738) and BIS/BAS (0.762 vs. 0.702) psychological models, whereas for the HEXACO model, images and videos exhibit the same acc = 0.754. The same observation is valid also for the F1 scores, where the video stimuli outperformed the images for all three models and overall. The superiority of the videos over the images can potentially be explained by the stronger affective nature of the former, which presumably evokes stronger emotional and physiological responses, facilitating an easier detection of the traits [12]. Although there were statistically significant differences between images and videos (Friedman test $p < .01$ for both accuracy and F1 scores), Wilcoxon post hoc tests indicated that the differences between the two types of stimuli were not significant, $p = .424$ for the accuracy and $p = .323$ for the F1 scores.

Analysing the traits individually, we observe that the video stimuli achieve a higher accuracy than the images for 10 traits, images outperform the videos for 5 traits, and for Openness they achieve the same accuracy. For the F1 scores, videos are superior to the images for 10 traits, while being inferior for 6. The better performance of the video stimuli is particularly pronounced for the BIS/BAS model, where they outperform the images for three out of four traits, both for accuracy and F1. Notably, the video stimuli accuracy outperform the images by more than 10% for 5 traits: Views, Narcissism, BAS Reward Responsiveness, Conscientiousness, and Honesty. However, the images outperform the videos by more than 10% for 4 traits: Tactics, BAS Fun Seeking, Agreeableness, and Extraversion. Thus, we conclude a slight better predictive performance when using video stimuli.

We also consider the combination of the image and video stimuli, shown in the "Both" columns. We highlight in bold the traits where the combined stimuli yielded accuracy or F1 scores superior to the best-performing individual type of stimuli, be it images or videos. When both types of stimuli are used, we observe acc = 0.860, which is 12.5% higher than for videos only and 17.0% higher than for images only. Also for the F1 scores, the combined F1 = 0.860 is 12.9% higher than for videos and 18.5% higher than for images only. In this case, the Wilcoxon post hoc tests showed that the combined image and video stimuli is significantly superior to both the image alone ($p < .001$ for accuracy and F1) and video alone ($p < .001$ for accuracy and F1). Hence, for the ETG signal, the performance of the video and image stimuli in combination is consistently higher than those of the individual types of stimuli for all three psychological models.

Analysing the individual traits, we observe that the accuracy of the combined stimuli outperforms the image and video stimuli individually for 11 traits and the F1 of the combined stimuli—for 12 traits. The accuracy of the combined stimuli is higher than the best-performing individual stimuli by more than 10% for 7 traits (Primary Psychopathy, Tactics, BIS, BAS Drive, BAS Fun Seeking, Resiliency, and Openness) and it is never inferior to the individual stimuli. For F1 scores, we observe comparable findings, with the combined stimuli being inferior to the best individual by 0.5% or less only for Secondary Psychopathy and Extraversion. The improvement in accuracy of the combined stimuli for D3, BIS/BAS, and HEXACO is 10.2%, 15.6%, and 12.6%, respectively, while for F1 it stands at 10.7%, 15.9%, and 13.1%, respectively.

In Table 10, we show the results of a similar comparison of image vs. video stimuli, but this time using the GSR signal. We observe that, like in Table 9, the video stimuli achieve a higher overall accuracy than images, 0.753 vs. 0.673. This result is observed consistently for the mean scores of the three psychological models: D3 (0.754 vs. 0.643), BIS/BAS (0.738 vs. 0.690), and HEXACO (0.762 vs. 0.690). This observation is also supported by the F1 scores, where the video stimuli steadily outperform the images for all three models and overall (0.752 vs 0.670). The Friedman test indicates there are statistically significant differences between the images and videos ($p < .01$ for both accuracy and F1 scores), but again the follow-up with the Wilcoxon post hoc tests (with $p < .0167$, Bonferroni correction for three tests) show no statistically significant difference between videos and images ($p = 0.037$ for accuracy and $p = 0.051$ for F1).

Table 10. Performance of the Image and Video Stimuli for the GSR Signal

	accuracy			F1 score		
Trait	**Image**	**Video**	**Both**	**Image**	**Video**	**Both**
Primary Psychopathy	0.571	**0.810**	**0.857**	0.559	**0.815**	**0.857**
Secondary Psychopathy	**0.667**	0.571	**0.714**	**0.677**	0.572	**0.727**
Tactics	0.762	**0.857**	**0.952**	0.758	**0.856**	**0.952**
Views	0.667	**0.810**	**0.857**	0.660	**0.807**	**0.853**
Morality	**0.762**	0.667	**0.762**	**0.769**	0.676	0.757
Narcissism	0.429	**0.810**	0.762	0.417	**0.810**	0.766
Mean D3	*0.643*	***0.754***	***0.817***	*0.640*	***0.756***	***0.819***
BIS	**0.810**	0.714	**0.905**	**0.810**	0.711	**0.904**
BAS Drive	**0.619**	0.571	**0.667**	**0.626**	0.552	**0.655**
BAS Fun Seeking	0.524	**0.762**	**0.905**	0.510	**0.762**	**0.904**
BAS Reward Responsiveness	0.810	**0.905**	**0.905**	0.800	**0.909**	0.903
Mean BIS/BAS	*0.690*	***0.738***	***0.845***	*0.687*	***0.733***	***0.842***
Agreeableness	0.667	**0.762**	**0.762**	0.668	**0.769**	**0.769**
Conscientiousness	0.762	**0.810**	**0.952**	0.752	**0.803**	**0.952**
Extraversion	**0.714**	**0.714**	**0.762**	**0.716**	**0.716**	**0.770**
Honesty	0.714	**0.810**	0.667	0.716	**0.804**	0.656
Resiliency	**0.667**	**0.667**	**0.810**	**0.668**	0.664	**0.811**
Openness	0.619	**0.810**	**0.857**	0.612	**0.811**	**0.857**
Mean HEXACO	*0.690*	***0.762***	***0.802***	*0.689*	***0.761***	***0.803***
Mean Overall	*0.673*	***0.753***	***0.818***	*0.670*	***0.752***	***0.819***

Breaking the model performance into individual traits, the accuracy of the video stimuli outperforms the images for 10 traits, the images are superior for 4 traits, and the accuracy for Extraversion and Resiliency is identical. A similar observation holds for F1, where the videos are superior for 10 traits, images—for 5 traits, and the same F1 is obtained for Extraversion. Notably, the video stimuli outperform the images for 4 out of the 6 D3 traits and for 4 out of the 6 HEXACO traits. The difference between the stimuli for the GSR signal is more pronounced than for ETG; for example, for GSR the accuracy of videos outperforms the images by 17.3% for D3 and 10.3% for HEXACO, compared to 5.4% difference and equal performance for the same models observed for the ETG signal. This explains the significant differences between the videos and images obtained for the GSR signal.

Considering the combination of the image and video stimuli, we obtain overall acc = 0.818, which is 8.7% higher than the best-performing video stimuli and 21.7% higher than the images. Similarly, the combined F1 = 0.819 of images and videos is 8.8% higher than F1 of the videos only, and 22.2% higher than the F1 of the images.This finding is statistically significant: the combined image and video stimuli is significantly superior to the video stimuli alone ($p = .014$ for accuracy and $p = .010$ for F1 scores) as well as to the image stimuli alone ($p < .001$ for both accuracy and F1 scores). These results reaffirm the ETG findings regarding the superiority of the combined video and image stimuli over either type of stimuli applied individually.

Analysing the individual traits, we observe that the accuracy of the combined stimuli outperforms the best-performing stimuli for 11 traits, while the combined accuracy actually drops only for 2 traits: Narcissism and Honesty. The dominance of the combined stimuli is comparable for the F1 scores, where it outperforms the individual stimuli for 11 traits. The combined outperforms the

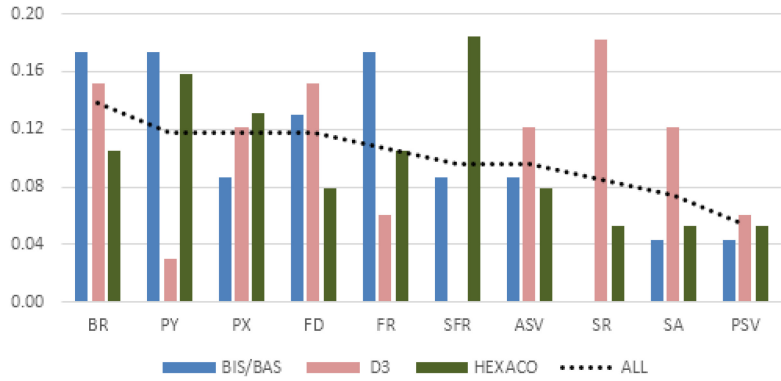

Fig. 5. Relative importance of features for each model and ETG signal.

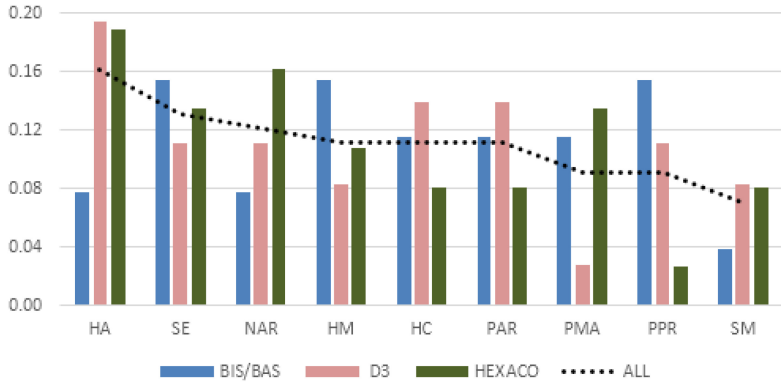

Fig. 6. Relative importance of features for each model and GSR signal.

best individual stimuli by more than 10% for 5 traits (Tactics, BIS, BAS Fun Seeking, Conscientiousness, and Resiliency) and this finding holds for both accuracy and F1. The improvement in accuracy of the combined stimuli for D3, BIS/BAS, and HEXACO is 8.4%, 14.5%, and 5.2%, respectively, while for F1 it is 8.3%, 14.8%, and 5.4%.

Hence, with respect to Q3 formulated at the outset of this section, we conclude that the videos are generally better stimuli than images, although the difference between the two depends on the signal being sensed and analysed. However, for both the ETG and GSR signals *the combined image and video stimuli yield better predictions* than either type of stimuli considered individually.

5.4 Predictive Stimuli and Features

We turn to Q4, which deals with the predictive value of the extracted features. For this, we calculate the normalised selection frequency for each extracted feature for the three personality models (D3, BIS/BAS, and HEXACO) separately. The rationale is that the more frequently selected features have more important predictive value; thus, the higher the value, the more important the feature is. Specifically, we counted how many times a feature was selected in a feature subset for a trait from each personality model and normalised this value per model. The results are shown in Figures 5 and 6 for the ETG and GSR signals, respectively.

Referring to the ETG data in Figure 5 and considering the detection of D3's traits, we note the dominance of the saccade rate (SR), which is supported by previous research that discovered

links between reduced saccade movements and facets of psychopathy [8]. Fixation duration (FD) and blink rate (BR) are the second most used features, which may be linked to the reduced saccade movements. This supports the work of Reference [58], which found that those with psychopathic traits tended to display unusual blink responses. The least selected feature is vertical pupil size (PY), in line with the findings of Reference [10], which showed no relations between the subjects' psychopathy scores and pupil diameter changes in response to affective stimuli. However, the trend is not supported by PX. No links to other D3 components were found in prior literature.

Looking into the BIS/BAS predictions, we highlight the importance of the BR feature. This aligns with the findings of Reference [31], which found significant correlations between the BAS scores and eye blink responses. In addition, we found that fixations (FR) are predictive of BIS/BAS traits, which is supported by previous research that linked the number and duration of fixations to BAS Drive and BAS Fun Seeking [64]. Vertical pupil dilation (PY) is also an important feature, as explained by the strong association between pupillary reactivity and both fear and anxiety [81], components of BIS.

Predictions of the HEXACO traits are dominated by the Saccade-Fixation Ratio (SFR), in line with Reference [64], which associated fixations with Extraversion, Agreeableness, and Neuroticism (inverse to Resiliency), although the individual FR and SR features are not selected often. Pupil size contributes to the second most selected features (both PY and PX), which is aligned with early works that studied traits such as Extraversion and Neuroticism (in this case, Resiliency) [22]. Summarising feature importance analysis for the ETG signal across the three psychological models, we note that the number of blinks and pupil size are the most important features.

Figure 6 shows the relative importance of the features for the three personality models using the GSR signal. We observe that the predictions of the D3 traits are strongly dominated by HA and HC, Hjorth's activity and complexity parameters. Hjorth's parameters were recently used in several works detecting stress and emotion [19, 53]. However, these used the EEG signal rather than GSR, such that this finding is important and novel. The other influential feature is PAR, which corresponds to the phasic area.

For the prediction of the HEXACO traits, the most important feature is again HA, followed by NAR. Again, to the best of our knowledge, no prior research links personality traits to Hjorth's parameters of the GSR signal, such that the dominance of the activity parameter is novel. However, NAR, which represents the number of sudomotor nerve activations, is a rather established factor for HEXACO traits. For example, skin conductance levels were found to correlate with Extraversion and Neuroticism (Big-5's counter-part of Resiliency) in Reference [9], while links to Agreeableness and Conscientiousness were established in Reference [7].

With respect to the BIS/BAS predictions, the three most frequently selected features are SE, HM, and PPR. Notably, SE communicates the energy of the GSR signal, whereas HM is the Hjorth's mobility frequency parameters, and PPR the phasic peaks rate. Hence, these three features collectively characterise the statistical properties of the GSR signal. To the best of our knowledge, no prior works looked at the links between BIS/BAS traits and measurement of GSR responses. Summarising the GSR analysis, we note the dominance of Hjorth's parameters, particularly activity that is the main predictor for both D3 and HEXACO.

Revisiting Q4, we conclude that *different features informed the predictions of different personality models*. For example, for the ETG signal, the saccade rate was most predictive of D3, blink rate, and pupil size—of BIS/BAS, and saccade-fixation rate—of HEXACO traits. Likewise, for GSR, Hjorth's activity was most predictive of D3 and HEXACO, while the signal energy, Hjorth's mobility, and the phasic peaks rates were most predictive of BIS/BAS. Overall, blinks and pupil size were found

Table 11. Performance of Classifiers Using the Combined Signal

Signal	Metric	AB	DT	LR	NB	RF	SVM	kNN
ETG or GSR individually	Accuracy	0.542	0.351	0.792	**0.860**	0.631	0.762	0.732
	F1 score	0.539	0.344	0.787	**0.860**	0.625	0.758	0.727
		(GSR)	(ETG)	(GSR)	**(ETG)**	(ETG)	(GSR)	(GSR)
ETG+GSR	Accuracy	0.539	0.313	0.872	**0.899**	0.673	0.830	0.830
	F1 score	0.535	0.308	0.871	**0.898**	0.668	0.828	0.827

to be the most predictive ETG features, while Hjorth's parameters were among the most predictive features for GSR.

5.5 Combining ETG and GSR

Having elaborately studied the performance of the ETG and GSR signals individually, we finally turn to Q5, which deals with the identification of the most accurate signal. For this analysis, we also consider a combination of the two signals. At the data collection stage, we were able to synchronise in time the signals captured by SMI glasses and Procomp Infiniti. As the temporal blocks referred to the same stimuli, we could use the combined set of ETG and GSR features for personality trait predictions. We re-run feature selection for the combined ETG+GSR feature set, re-train the classifiers, and conduct again the analyses presented in Sections 5.1–5.4.

We first revisit Q1 to validate the dominance of NB for the combined signal. Table 11 shows the accuracy and F1 of the seven classifiers using the ETG+GSR signals and compares them with the best-performing classifier using either of the two signals individually (extracted from Table 6 and the name of the classifier is given). It can be observed that the combined ETG+GSR signal is superior to the best-performing individual signal for most classifiers, both for the accuracy and F1 scores. For example, considering the accuracy of the three top-performing classifiers (NB, LR, and SVM), we note that ETG+GSR beats the best-performing individual signal, by 4.5%, 10.2%, and 9.0%, respectively. A similar dominance of the combined ETG+GSR signal is obtained also for F1.

More importantly, NB outperforms all other classifiers also for the combined signal. Namely, NB outperforms the second-best LR classifier by 3.1% for accuracy and by 3.2% for F1, and the third-best SVM by 8.2% and 8.5%, respectively. The superiority of NB is statistically significant compared to all other classifiers for accuracy and F1 scores (both Friedman $p < .001$). The post hoc Wilcoxon tests show NB is statistically superior to all other classifiers ($p < .0071$, Bonferroni correction for seven tests) except for LR ($p = .124$ for accuracy and $p = .068$ for F1 score). Hence, these results re-affirm our earlier finding that NB is the most appropriate machine learning method out of the seven studied classifiers, so in the following analyses, we focus again only on NB.

Q3 clearly demonstrated that the combination of image and video stimuli is superior to either type of stimuli applied in isolation. To this end, we now revisit Q2 addressing the individual personality traits for the combined ETG+GSR signal using both types of stimuli. To this end, Table 12 shows the accuracy and F1 scores obtained for every trait by the ETG and GSR signals and by the combined ETG+GSR signal. We highlight in bold again the best-performing individual signal as well as the cases when ETG+GSR is superior to the best individual signal.

Comparing the accuracy scores obtained by the ETG and GSR signals, we note the superiority of the former. Specifically, ETG obtains a higher accuracy for 10 traits, GSR—for 4 traits, and for BIS and Openness their accuracy is identical. The superiority of ETG over GSR comes through more clearly considering the mean accuracy scores across the three psychological models. Here, ETG achieves a higher accuracy for all the models and is better than GSR by 4.9% for D3 traits, by 4.2% for BIS/BAS traits, and by 5.9% for HEXACO traits. Naturally, ETG also achieves a higher overall

Table 12. Performance of the ETG, GSR, and ETG+GSR Signals for the 16 Traits

	accuracy			F1 score		
Trait	**ETG**	**GSR**	**ETG+GSR**	**ETG**	**GSR**	**ETG+GSR**
Primary Psychopathy	0.762	**0.857**	**0.905**	0.756	**0.857**	**0.901**
Secondary Psychopathy	**0.857**	0.714	**0.857**	**0.856**	0.727	**0.856**
Tactics	0.905	**0.952**	**0.952**	0.905	**0.952**	**0.952**
Views	**0.905**	0.857	**0.952**	**0.904**	0.853	**0.953**
Morality	**0.905**	0.762	**0.905**	**0.906**	0.757	**0.906**
Narcissism	**0.810**	0.762	**0.857**	**0.812**	0.766	**0.861**
Mean D3	***0.857***	*0.817*	***0.905***	***0.856***	*0.819*	***0.905***
BIS	**0.905**	**0.905**	**0.952**	**0.904**	**0.904**	**0.952**
BAS Drive	**0.810**	0.667	**0.905**	**0.819**	0.655	**0.904**
BAS Fun Seeking	**0.952**	0.905	**0.952**	**0.952**	0.904	**0.952**
BAS Reward Responsiveness	0.857	**0.905**	**0.952**	0.850	**0.903**	**0.953**
Mean BIS/BAS	***0.881***	*0.845*	***0.940***	***0.882***	*0.842*	***0.940***
Agreeableness	**0.905**	0.762	0.667	**0.905**	0.769	0.656
Conscientiousness	0.810	**0.952**	0.857	0.810	**0.952**	0.860
Extraversion	**0.810**	0.762	**0.857**	**0.811**	0.770	**0.858**
Honesty	**0.810**	0.667	**0.905**	**0.810**	0.656	**0.901**
Resiliency	**0.905**	0.810	**1.000**	**0.905**	0.811	**1.000**
Openness	**0.857**	**0.857**	**0.905**	**0.858**	0.857	**0.905**
Mean HEXACO	***0.849***	*0.802*	***0.865***	***0.850***	*0.803*	***0.863***
Mean Overall	***0.860***	*0.818*	***0.899***	***0.860***	*0.819*	***0.898***

accuracy than GSR, 0.860 vs. 0.818, which amounts to a 5.1% improvement. There is a significant difference in accuracy between ETG, GSR, and ETG+GSR (Friedman $p = .001$). All pair differences are statistically significant except the ETG's superiority over GSR (post hoc Wilcoxon $p = .104$).

The F1 scores generally exhibit the same trends as accuracy. Namely, ETG achieves higher F1 scores than GSR for 11 traits, for 4 traits GSR is better, and for BIS the 2 are similar. The overall F1 of ETG is again 5.1% better than that of GSR and also all three model-specific mean F1 scores of ETG are higher: by 4.6% for D3, by 4.7% for BIS/BAS, and by 5.9% for HEXACO. Similarly, there is a significant difference in F1 scores between ETG, GSR, and ETG+GSR (Friedman $p = .001$), but the post hoc Wilcoxon signed-rank test shows no significant difference in F1 scores between ETG and GSR when considered individually ($p = .117$). These results contrast previous conclusions of References [89] and [39], where GSR was found to outperform ETG. We posit that the differences should be attributed to the features being extracted and the nature of the driving-related tasks used in those works, which differ substantially from personality trait detection and affective image and video stimuli deployed in our work.

Turning to the performance of ETG+GSR versus the individual signals, we note that the combined signal outperforms its individual components. Specifically, ETG+GSR achieves accuracy scores higher than the best of the two signals individually for 10 traits out of the 16. The combined signal is particularly dominant for the D3 and BIS/BAS models, where it outperforms or equals individual signals for all traits. Interestingly, only for Agreeableness and Conscientiousness ETG+GSR is beaten. ETG+GSR outperforms the better-performing ETG signal by 5.6% for D3 predictions, 6.8% for BIS/BAS, and 1.9% for HEXACO. Overall, ETG+GSR achieves acc = 0.899, which is 4.5% better than ETG and 9.8% better than GSR individually. The dominance in accuracy

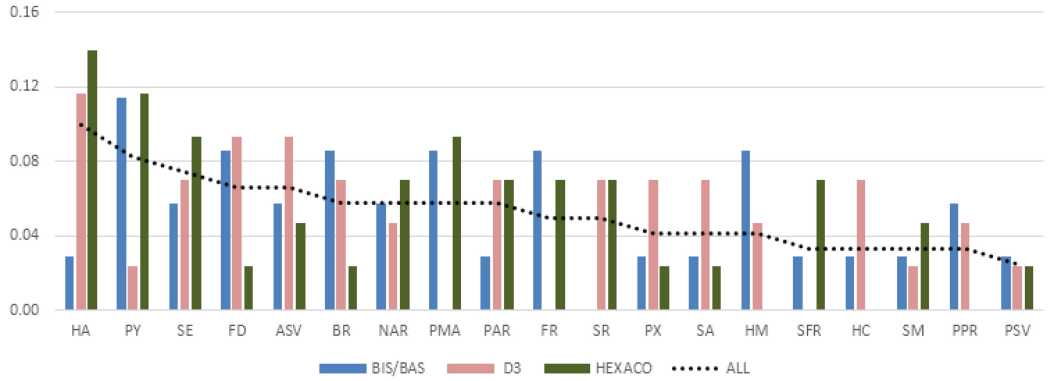

Fig. 7. Relative importance of features for each model and ETG+GSR signal.

of ETG+GSR over its individual components is statistically significant for GSR (post hoc Wilcoxon $p = .009$) and borderline for ETG ($p = .017$).

These findings are also supported by the F1 scores achieved by the combined signal. ETG+GSR achieves F1 = 0.898, which is 4.4% higher than F1 = 0.860 of the ETG signal and 9.7% higher than F1 = 0.819 of GSR considered individually. This is also evident for the three psychological models, where ETG+GSR is consistently better than the best-performing ETG signal (by 5.7%, 6.7%, and 1.6%, respectively). The dominance of ETG+GSR is also clearly evident when analysing individual traits. It is the best-performing signal for four out of the six D3 traits, for three out of the four BIS/BAS traits, and for four out of the six HEXACO traits. Again, it is inferior to the individual signals for two HEXACO traits of Agreeableness and Conscientiousness. The F1 scores of ETG+GSR and ETG are similar to the best-performing individual signal in all other cases. The F1 trend of ETG+GSR outperforming the individual stimuli is statistically significant for both GSR ($p = .012$) and ETG ($p = .010$).

Finally, we turn to the analysis of most important features in the combined ETG+GSR feature set. Figure 7 shows the relative importance of the features for predictions of the three models. Overall, we observe a comparable contribution of the two signals. For example, in the list of top five features there are three GSR features and two ETG features, while in top twelve each signal contributes six features. Note that there are some fluctuations in the ranking of features. For example, BR, top-performing ETG features ranked only sixth in the combined list, while HA, the top-two GSR feature, retains its position in the combined list. Similarly, the second most important GSR feature, SE, stays in the top three for combined features. We posit that these fluctuations are explained by dependencies between the features across the two signals that are picked by the CFS feature selection.

Focussing on the three psychological models, we highlight the strong dominance of PY and other ETG features in the predictions of BIS/BAS traits; PY also being the second most important feature across all the models. While HA is one of the lowest performing for BIS/BAS, it is clearly the best predictive feature for both D3 and HEXACO. D3 is then well predicted by FD and ASV while PY is a strong second predictor for HEXACO. It is difficult to position these findings in the context of prior research, as no previous work looked at combination of these two signals and also extracted the same feature set.

Hence, summarising Q5 outlined at the beginning of the session, we conclude that the ETG signal allows generating slightly more accurate trait predictions than GSR, although the difference between the two is not significant. However, a *combination of ETG and GSR further improves the*

predictions over either of the two signals considered individually, and features from both signals inform the trait predictions.

6 DISCUSSION

In this work, we developed a framework for predicting human personality traits using physiological responses to external stimuli. We found that the Naive Bayes algorithm, in conjunction with feature selection, substantially outperformed other machine learning algorithms. Overall, seven traits were predicted with the ETG signal and five traits with the GSR signal, all with accuracy levels greater than 0.9. Comparing the image and video stimuli, we found that the latter were more predictive than the former for both the signals, while the combined use of images and videos resulted in the most accurate predictions. Predictive models using ETG were slightly more accurate than those using GSR. However, the combination of the ETG and GSR signals achieved the best result: an overall mean accuracy of 0.899 and mean F1-score of 0.898 across the 16 traits, which is a notably high result, considering the intricate nature of predicting human personality traits.

It is important to highlight that the evaluated instantiation of the proposed framework did not involve any custom-made components. That is, the deployed eye-tracking and skin conductivity sensors were commercial-grade, the psychological tools were established and well-studied models and inventories, and all the image and video stimuli were pre-existing. Furthermore, the feature extraction and classification algorithms are widely used machine learning tools that have previously been deployed in numerous applications and are not specific to physiological signal processing or personality detection.

6.1 Comparison

We would like to analytically compare our results to the closest line of research on personality detection using physiological signals [1, 18, 72, 78]. Our findings largely align with the main observations made in those papers. However, our work additionally shows several notable advantages of a significant practical benefit:

- *Personality traits.* Previous works focused on the predictions of the Big-5 traits only. In our work, not only do we use the more recent HEXACO model that introduces an additional trait of Honesty to the Big-5, but we also complement this with traits from the D3 and BIS/BAS models. Altogether, our method is capable of predicting more than three times the number of traits predicted in References [1, 18, 72, 78]. These offer an encompassing perspective on human personality and may be useful in practical scenarios, like hiring decisions, particularly for the jobs that would require screening out people scoring high or low on a particular trait.
- *Classification accuracy.* Compared to previous research, our method achieves substantially higher classification accuracy and F1 scores. Specifically, References [1, 18, 72, 78] conducted a two-class classification, whereas our work addresses a three-class classification. Thus, our random-guess baselines of 33% compared to the 50% baseline of previous works. The F1-scores reported in References [1, 18, 72, 78] generally hover between the 0.5 and 0.8 marks, while our results achieve accuracy levels as high as 1.0, with the mean accuracy of 0.899 and mean F1 of 0.898 across the 16 traits being predicted. Hence, previous work achieved a 30%–40% improvement, while we achieve 170% improvement over the random baseline.
- *Duration of stimuli.* Previous works [1, 18, 72, 78] required the subjects to be exposed to the stimuli for substantially longer periods of time. For example, References [78] and [72] used 36 video clips, on average 80 seconds long each, which brought the overall duration of the video stimuli to 48 minutes. In the more recent work, they used four longer videos

that summed up to 85 minutes [72]. In our experimental setting, the image and video stimuli required 9.5 and 14 minutes, respectively, which keeps the combined duration of the stimuli under 25 minutes—much shorter than the above times. Also, reasonably high levels of classification accuracy were achieved with one type of stimuli only (either images or videos), which would require even less time.

— *Deployed sensors.* In our work, we deployed the ETG and GSR sensors, while previous works [1, 18, 72, 78] used a substantially larger range of sensing technologies. Specifically, in all four papers, the authors used the EEG, GSR sensors, ECG (heart rate), as well as video-based face feature trackers. These sensing technologies are usually more complex, sometimes more expensive, usually more obtrusive than the ones we focused on with the present framework. The high performance we achieved indicates that, in practice, fewer sensors coupled with an advanced feature extraction and state-of-the-art classifiers may be sufficient. While ETG+GSR produces sensibly better results than individual sensors, these latter could still be used in isolation and produce reasonably high accuracy. This should allow practitioners or interactive system designers to address practical constraints by using a single sensor, yet obtain reliable enough classification performance.

6.2 Limitations

While our results are promising, there are several limitations that require attention and need to be addressed in follow-up works.

— The first limitation refers to the reasonably small sample size. Although the leave-one-out validation with 21 subjects produced solid results, more subjects should be recruited to better understand the results, validate our findings, and replicate them for other sensors and traits.

— The second limitation refers to the subject recruitment, not based on any psychological or clinical criteria. Thus, we were unlikely to have subjects on the extreme ends of the scales for some traits, especially the D3 traits, also evident from the range of personality scores in Table 5. Due to the equal-frequency discretisation of subjects, our results are likely to demonstrate reliable predictive accuracy for a population of normative subjects with medium trait values, while a targeted recruitment would be required for validation in the extreme ranges of the traits.

— The third limitation is the use of the equal-frequency binning of the subjects, not based on norms or psychological theories. Given the second limitation, replication on a larger sample, using norms, whenever these are available, is needed to determine the generalisability of our findings. Future research should also employ a full spread of scores on psychological traits of interest to increase the authenticity of results within predictive models.

— The last limitation refers to the "off-the-shelf" nature of the stimuli, sensors, feature extraction, and classifiers. While this can be interpreted as a limitation, e.g., with regards to the signal quality and predictive accuracy, it is also a door-opener for future improvements and a strength. Although we managed to accurately detect personality with these off-the-shelf components, we posit that accuracy can be substantially improved by tailoring the components of the framework to the trait prediction task, e.g., by tuning the classifier parameters.

7 CONCLUSIONS AND FUTURE WORK

In this article, we consider the task of objective detection of personality traits using physiological responses to external stimuli. Specifically, we propose a framework, which combines external stimuli that trigger physiological responses, sensing technologies that capture these responses,

and machine learning methods that detect the personality traits based on the responses. To obtain the class labels and train the machine learning classifiers, we deploy personality inventories. The trained classifiers can then accurately detect personality traits for new subjects. We evaluate a specific instantiation of the framework, which uses affective image and video stimuli and two types of physiological responses to these stimuli: eye activity and skin conductance. Our work demonstrates that personality traits can be accurately detected, suggesting possible use in practical applications to supplement the traditional forms of assessment or to provide a practical alternative for tailored human-computer interaction through content personalisation or the design of user-aware interactive intelligent systems.

Revisiting the research questions, we established that: (i) Naive Bayes was the most accurate classification method; (ii) traits associated with affect were predicted more accurately than traits associated with behaviour and cognition for both signals; (iii) video stimuli were more predictive than the images, while the best predictions were obtained by combining the two types of stimuli; (iv) predictive features differed across the models, consistently with previous psychology research; and (v) the use of the ETG signal resulted in slightly more accurate prediction models than GSR, while the combination of the two signals achieved the highest accuracy. Our findings enhanced prior research by considering a broader range of traits and models, and improving the trait prediction accuracy, while deploying only commercial-grade sensing technologies and reducing the data acquisition times.

Future research should address the identified limitations, including experimenting with a larger cohort of participants and in different scenarios, such as gaming, driving, and more. For psychology practitioners and clinicians, it will be important to validate our method with populations having an established pathology [8]. In addition to the eye-tracking and skin conductance signals, other physiological signals such as EEG should be investigated and may further improve the results. Decisions on the sensors to deploy in practice should be based on the level of obtrusiveness, cost, complexity to setup and calibration, and desired accuracy. The individual components of our method (stimuli, sensors, feature extraction, classifiers) may also be refined in the future. While we managed to establish high levels of accuracy using off-the-shelf components, the performance may be further improved by tailoring the components of the framework to the specific trait prediction task.

Another stream of work that may potentially improve the accuracy and reliability of the detected personality traits refers to the use of behavioural and interaction data. While physiological responses to stimuli were shown to achieve accurate predictions, the availability of accurate and affordable sensing technologies is still reasonably limited, which may hinder wide deployment of the proposed framework. On the contrary, behavioural data (e.g., web browsing or multimedia consumption logs) and interaction data (e.g., voice conversations with a smart assistant) are substantially easier to capture and more abundant. This brings to the fore the question of establishing personality traits using such data, which has been studied on social media [42, 61] and in written essays [49]. We posit that enriching physiological responses with easily obtainable behavioural and interaction data has the potential to further improve the accuracy of the predictions.

The presented feature importance analysis concentrated on the predictive power of various features for various personality models. However, the granularity level of such an analysis can be improved on both dimensions being analysed. Considering the extracted feature, the analysis can potentially focus on the features extracted for specific stimuli, be it images or videos at large, or even the tenderness clip or the disgust clip. Likewise, on the personality dimension the analysis can drill down into individual traits being predicted and surface the importance of low-level stimuli-specific features for predictions of particular traits. Beyond the mere feature importance information, this fine-grained analysis can uncover valuable information on the stimuli and sensing

technologies needed for detection of individual personality traits. For example, it can yield specific guidelines regarding the shortest set of stimuli and cheapest sensing technologies that will allow detection of a target trait with the desired level of accuracy.

An additional important area calling for future research is the thorough investigation of how each scenario and type of stimuli influences personality detection, as certain scenarios and stimuli can be linked to certain traits stronger than to others. Hence, it is important to deploy the right stimuli to evoke responses that are predictive of the target trait. Also, the contribution of various physiological features to personality detection should be further studied. In this way, the neuro-psychological association between features and personality can be investigated and better understood. Finally, we would like to study physiological responses beyond personality detection tasks, e.g., for gauging the effect of social media on users. These exciting veins of future work may evolve into broader cross-disciplinary research initiatives combining capabilities from human-computer interaction, psychology, sensing technologies, signal processing, machine learning, and physiology.

REFERENCES

[1] Mojtaba K. Abadi, Juan A. Miranda-Correa, Julia Wache, Heng Yang, Ioannis Patras, and Nicu Sebe. 2015. Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG'15)*. 1–8.

[2] Daniel R. Ames, Paul Rose, and Cameron P. Anderson. 2006. The NPI-16 as a short measure of narcissism. *J. Res. Person.* 40, 4 (2006), 440–450.

[3] Jeromy Anglim, Stefan Bozic, Jonathon Little, and Filip Lievens. 2018. Response distortion on personality tests in applicants: Comparing high-stakes to low-stakes medical settings. *Adv. Health Sci. Educ.* 23 (10 2018), 311–321.

[4] Michael C. Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E. De Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. 2004. A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *J. Person. Social Psychol.* 86, 2 (2004), 356.

[5] Shlomo Berkovsky, Tsvi Kuflik, and Francesco Ricci. 2012. The impact of data obfuscation on the accuracy of collaborative filtering. *Expert Syst. Appl.* 39, 5 (2012), 5033–5042.

[6] Shlomo Berkovsky, Ronnie Taib, Irena Koprinska, Eileen Wang, Yucheng Zeng, Jingjie Li, and Sabina Kleitman. 2019. Detecting personality traits using eye-tracking data. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'19)*. 221.

[7] Erdal Binboga, Senol Guven, Fatih Çatıkkaş, Onur Bayazıt, and Serdar Tok. 2012. Psychophysiological responses to competition and the big five personality traits. *J. Hum. Kinet.* 33 (2012), 187–194.

[8] Sabrina Boll and Matthias Gamer. 2016. Psychopathic traits affect the visual exploration of facial expressions. *Biolog. Psychol.* 117 (2016), 194–201.

[9] Claudia Chloe Brumbaugh, Ravi Kothuri, Carl Marci, Caleb Siefert, and Donald D. Pfaff. 2013. Physiological correlates of the Big 5: Autonomic responses to video presentations. *Appl. Psychophys. Biofeed.* 38, 4 (2013), 293–301.

[10] Daniel T. Burley, Nicola S. Gray, and Robert J. Snowden. 2017. As far as the eye can see: Relationship between psychopathic traits and pupil response to affective stimuli. *PloS One* 12, 1 (2017), 1–22.

[11] James N. Butcher, Mera M. Atlis, and Jungwon Hahn. 2004. The Minnesota Multiphasic Personality Inventory–2 (MMPI-2). In *Comprehensive Handbook of Psychological Assessment: Personality Assessment*. John Wiley & Sons Inc., 30–38.

[12] Suzanne R. Byrnes. 1996. The effect of audio, video, and paired audio-video stimuli on the experience of stress. *J. Mus. Ther.* 33, 4 (1996), 248–260.

[13] Iván Cantador, Ignacio Fernández-Tobías, and Alejandro Bellogín. 2013. Relating personality types with user preferences in multiple entertainment domains. In *Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE'13)*. 13–28. http://ceur-ws.org/Vol-997/.

[14] Charles S. Carver and Teri L. White. 1994. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *J. Person. Social Psychol.* 67, 2 (1994), 319.

[15] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M. Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2013. Multimodal behavior and interaction as indicators of cognitive load. *ACM Trans. Interact. Intell. Syst.* 2, 4 (2013), 22:1–22:36.

[16] C. Robert Cloninger, Dragan M. Svrakic, and Thomas R. Przybeck. 1998. A psychobiological model of temperament and character. *Dev. Psychiat. Its Complex.* 50, 12 (1998), 1–16.

[17] Hope R. Conte and Robert Plutchik. 1981. A circumplex model for interpersonal personality traits. *J. Person. Social Psychol.* 40, 4 (1981), 701.

[18] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. 2018. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* (2018). DOI: 10.1109/TAFFC.2018.2884461

[19] Deepan Das, Tanuka Bhattacharjee, Shreyasi Datta, Anirban Dutta Choudhury, Pratyusha Das, and Arpan Pal. 2017. Classification and quantitative estimation of cognitive stress from in-game keystroke analysis using EEG and GSR. In *Proceedings of the IEEE Life Sciences Conference (LSC'17)*. 286–291.

[20] Ariel L. Del Gaizo and Diana M. Falkenbach. 2008. Primary and secondary psychopathic-traits and their relationship to perception and experience of emotion. *Person. Indiv. Dif.* 45, 3 (2008), 206–212.

[21] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual differences in gaze patterns for web search. In *Proceedings of the Symposium on Information Interaction in Context*. 185–194.

[22] Hans J. Eysenck. 1983. Psychophysiology and personality: Extraversion, neuroticism and psychoticism. In *Individual Differences and Psychopathology (Physiological Correlates of Human Behaviour, Vol. 3)*. Academic Press, 13–30.

[23] Gerry Fahey. 2018. Faking good and personality assessments of job applicants: A review of the literature. *DBS Bus. Rev.* 2 (2018), 45–68.

[24] Arik Friedman, Shlomo Berkovsky, and Mohamed Ali Kâafar. 2016. A differential privacy framework for matrix factorization recommender systems. *User Model. User-adapt. Interact.* 26, 5 (2016), 425–458.

[25] Mihai Gavrilescu and Nicolae Vizireanu. 2017. Predicting the sixteen personality factors (16PF) of an individual by analyzing facial features. *EURASIP J. Image Video Proc.* 2017, 1 (2017), 59.

[26] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from Twitter. In *Proceedings of the International Conference on Social Computing (SocialCom'11)*. 149–156.

[27] Harrison G. Gough. 1960. The adjective check list as a personality assessment research technique. *Psychol. Rep.* 6, 1 (1960), 107–122.

[28] Jeffrey A. Gray. 1990. Brain systems that mediate both emotion and cognition. *Cog. Emot.* 4, 3 (1990), 269–288.

[29] Alberto Greco, Antonio Lanata, Gaetano Valenza, Enzo Pasquale Scilingo, and Luca Citi. 2014. Electrodermal activity processing: A convex optimization approach. In *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'14)*. IEEE, 2290–2293.

[30] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2016. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng.* 63, 4 (2016), 797–804.

[31] Daniel F. Gros. 2011. Startle inhibition to positive-activated compared to neutral stimuli: Variations in self-reported behavioral approach. *J. Psychopath. Behav. Assess.* 33, 3 (2011), 308–314.

[32] Mark A. Hall. 2000. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the 17th International Conference on Machine Learning*. 359–366.

[33] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explor.* 11, 1 (2009), 10–18.

[34] Bo Hjorth. 1970. EEG analysis based on time domain properties. *Electroenceph. Clin. Neurophys.* 29, 3 (1970), 306–310.

[35] Sabrina Hoppe, Tobias Loetscher, Stephanie A. Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Front. Hum. Neurosci.* 12 (2018), 105.

[36] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Appl. Logist. Regr.* Vol. 398. John Wiley & Sons.

[37] Peter K. Jonason and Laura Krause. 2013. The emotional deficits associated with the Dark Triad traits: Cognitive empathy, affective empathy and alexithymia. *Person. Indiv. Diff.* 55, 5 (2013), 532–537.

[38] Anthony F. Jorm, Helen Christensen, Alexander S. Henderson, Patricia A. Jacomb, Ailsa E. Korten, and Bryan Rodgers. 1998. Using the BIS/BAS scales to measure behavioural inhibition and behavioural activation: Factor structure, validity and norms in a large community sample. *Person. Indiv. Diff.* 26, 1 (1998), 49–58.

[39] Enkelejda Kasneci, Thomas Kübler, Klaus Broelemann, and Gjergji Kasneci. 2017. Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth. *Comput. Hum. Behav.* 68 (2017), 450–455.

[40] Stéphanie Khalfa, Peretz Isabelle, Blondin Jean-Pierre, and Robert Manon. 2002. Event-related skin conductance responses to musical emotions in humans. *Neurosci. Lett.* 328, 2 (2002), 145–149.

[41] Ron Kohavi and Dan Sommerfield. 1995. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *Proceedings of the 36th Annual International Conference on Knowledge Discovery and Data Mining*. 192–197.

[42] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Nat. Acad. Sci.* 110, 15 (2013), 5802–5805.

[43] Nicole C. Krämer and Stephan Winter. 2008. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *J. Media Psychol.* 20, 3 (2008), 106–116.

[44] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. 2008. *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual.* Technical Report A-8. University of Florida, Gainesville, FL.

[45] Michael R. Levenson, Kent A. Kiehl, and Cory M. Fitzpatrick. 1995. Assessing psychopathic attributes in a noninstitutionalized population. *J. Person. Social Psychol.* 68, 1 (1995), 151.
[46] Alexandros Liapis, Christos Katsanos, Dimitris Sotiropoulos, Michalis Xenos, and Nikos Karousos. 2015. Recognizing emotions in human computer interaction: Studying stress using skin conductance. In *Human-computer Interaction*. Springer, 255–262.
[47] William L. Libby and Donna Yaklevich. 1973. Personality determinants of eye contact and direction of gaze aversion. *J. Person. Social Psychol.* 27, 2 (1973), 197.
[48] Kai K. Lim, Max Friedrich, Jenni Radun, and Kristiina Jokinen. 2013. Lying through the eyes: Detecting lies through eye movements. In *Proceedings of the Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction*. 51–56.
[49] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intell. Syst.* 32, 2 (2017), 74–79.
[50] Gerald Matthews, Ian J. Deary, and Martha C. Whiteman. 2003. *Personality Traits*. Cambridge University Press.
[51] Robert R. McCrae and Paul T. Costa Jr. 1999. A five-factor theory of personality. *Handb. Person.: Theor. Res.* 2 (1999), 139–153.
[52] Bruce S. McEwen and Robert M. Sapolsky. 1995. Stress and cognitive function. *Curr. Opin. Neurobiol.* 5, 2 (1995), 205–216.
[53] Raja Majid Mehmood and Hyo Jong Lee. 2015. EEG based emotion recognition from human brain using Hjorth parameters and SVM. *Int. J. Bio-Sci. Bio-technol.* 7, 3 (2015), 23–32.
[54] Lyle H. Miller and Barry M. Shmavonian. 1965. Replicability of two GSR indices as a function of stress and cognitive activity. *J. Person. Social Psychol.* 2, 5 (1965), 753.
[55] Leslie C. Morey. 2015. *Personality Assessment Inventory (PAI)*. Wiley Online Library.
[56] Carolyn C. Morf and Frederick Rhodewalt. 2001. Unraveling the paradoxes of narcissism: A dynamic self-regulatory processing model. *Psychol. Inq.* 12, 4 (2001), 177–196.
[57] Tim Morris, Paul Blenkhorn, and Farhan Zaidi. 2002. Blink detection for real-time eye tracking. *J. Netw. Comput. Applic.* 25, 2 (2002), 129–143.
[58] Christopher J. Patrick, Margaret M. Bradley, and Peter J. Lang. 1993. Emotion in the criminal psychopath: Startle reflex modulation. *J. Abnorm. Psychol.* 102, 1 (1993), 82–92.
[59] Delroy L. Paulhus and Kevin M. Williams. 2002. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *J. Res. Person.* 36, 6 (2002), 556–563.
[60] Lane Phillips, Victoria Interrante, Michael Kaeding, Brian Ries, and Lee Anderson. 2012. Correlations between physiological response, gait, personality, and presence in immersive virtual environments. *Presence* 21, 2 (2012), 119–141.
[61] Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. You are what you tweet: Personality expression and perception on Twitter. *J. Res. Person.* 46, 6 (2012), 710–718.
[62] George E. Raptis, Christos A. Fidas, and Nikolaos M. Avouris. 2017. On implicit elicitation of cognitive strategies using gaze transition entropies in pattern recognition tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1993–2000.
[63] John F. Rauthmann. 2013. Investigating the MACH–IV with item response theory and proposing the trimmed MACH. *J. Person. Assess.* 95, 4 (2013), 388–397.
[64] John F. Rauthmann, Christian T. Seubert, Pierre Sachse, and Marco R. Furtner. 2012. Eyes as windows to the soul: Gazing behavior is related to personality. *J. Res. Person.* 46, 2 (2012), 147–156.
[65] John F. Rauthmann and Theresa Will. 2011. Proposing a multidimensional Machiavellianism conceptualization. *Social Behav. Person.: Int. J.* 39, 3 (2011), 391–403.
[66] Richard M. Ryckman. 2012. *Theories of Personality*. Cengage Learning.
[67] Gerard Saucier. 2009. Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *J. Person.* 77, 5 (2009), 1577–1614.
[68] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cog. Emot.* 24, 7 (2010), 1153–1172.
[69] Louis L. Scharf and Cédric Demeure. 1991. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Vol. 63. Addison-Wesley Reading.
[70] Dara Schniederjans, Edita S. Cao, and Marc Schniederjans. 2013. Enhancing financial performance with social media: An impression management perspective. *Dec. Supp. Syst.* 55, 4 (2013), 911–918.
[71] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Proceedings of the CHI'07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2651–2656.
[72] Ramanathan Subramanian, Julia Wache, Mojtaba Abadi, Radu Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Trans. Affect. Comput.* 9, 2 (2016), 147–160.

[73] Robert P. Tett, Jenna R. Fitzke, Patrick L. Wadlington, Scott A. Davies, Michael G. Anderson, and Jeff Foster. 2009. The use of personality test norms in work settings: Effects of sample size and relevance. *J. Occup. Organiz. Psychol.* 82, 3 (2009), 639–659.

[74] Isabel Thielmann, Benjamin E. Hilbig, Ingo Zettler, and Morten Moshagen. 2016. On measuring the sixth basic personality dimension: A comparison between HEXACO honesty-humility and Big Six honesty-propriety. *Assessment* 24, 8 (2016), 1024–1036.

[75] Amit Tiroshi, Shlomo Berkovsky, Mohamed Ali Kaafar, Terence Chen, and Tsvi Kuflik. 2013. Cross social networks interests predictions based ongraph features. In *Proceedings of the ACM Conference on Recommender Systems*. 319–322.

[76] David Vallet, Shlomo Berkovsky, Sebastien Ardon, Anirban Mahanti, and Mohamed Ali Kâafar. 2015. Characterizing and predicting viral-and-popular video content. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 1591–1600.

[77] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Trans. Affect. Comput.* 5, 3 (2014), 273–291.

[78] Julia Wache, Ramanathan Subramanian, Mojtaba Khomami Abadi, Radu-Laurentiu Vieriu, Nicu Sebe, and Stefan Winkler. 2015. Implicit user-centric personality recognition based on physiological responses to emotional videos. In *Proceedings of the International Conference on Multimodal Interaction*. 239–246.

[79] Michael Wai and Niko Tiliopoulos. 2012. The affective and cognitive empathic nature of the dark triad of personality. *Person. Indiv. Diff.* 52, 7 (2012), 794–799.

[80] Peter Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* 15, 2 (1967), 70–73.

[81] Tara L. White and Richard A. Depue. 1999. Differential association of traits of fear and anxiety with norepinephrine- and dark-induced pupil reactivity. *J. Person. Social Psychol.* 77, 4 (1999), 863–877.

[82] Anne-Kathrin Wilbers, Alina Vennekoetter, Moritz Kuster, Kai-Christoph Hamborg, and Kai Kaspar. 2015. Personality traits and eye movements: An eye-tracking and pupillometry study. In *Proceedings of the European Conference on Eye Movements*. 269.

[83] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

[84] Jie Xu, Yang Wang, Fang Chen, and Eric Choi. 2011. Pupillary response based cognitive workload measurement under luminance changes. In *Proceedings of the IFIP Conference on Human-computer Interaction*. 178–185.

[85] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D. Abowd, and James M. Rehg. 2012. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the ACM Conference on Ubiquitous Computing*. 699–704.

[86] Lei Yu and Huan Liu. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the International Conference on Machine Learning (ICML'03)*. 856–863.

[87] Qifei Zhang, Xiangwei Lai, and Guangyuan Liu. 2016. Emotion recognition of GSR based on an improved quantum neural network. In *Proceedings of the 8th International Conference on Intelligent Human-machine Systems and Cybernetics (IHMSC'16)*, Vol. 1. IEEE, 488–492.

[88] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. 2018. Personality-aware personalized emotion recognition from physiological signals. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'18)*. 1660–1667.

[89] Jianlong Zhou, Jinjun Sun, Fang Chen, Yang Wang, Ronnie Taib, Ahmad Khawaji, and Zhidong Li. 2015. Measurable decision making with GSR and pupillary analysis for intelligent user interface. *ACM Trans. Comput.-human Interact.* 21, 6 (2015), 33.

[90] Michael J. Zickar and Fritz Drasgow. 1996. Detecting faking on a personality instrument using appropriateness measurement. *Appl. Psychol. Meas.* 20, 1 (1996), 71–87.

[91] Matthias Ziegler, Carolyn MacCann, and Richard Roberts. 2011. *New Perspectives on Faking in Personality Assessment*. Oxford University Press.

[92] Lisa M. Pytlik Zillig, Scott H. Hemenover, and Richard A. Dienstbier. 2002. 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. *Person. Social Psychol. Bull.* 28, 6 (2002), 847–858.

Received February 2019; revised December 2019; accepted February 2020