Performance and Usability Evaluation of Brainwave Authentication Techniques with Consumer Devices

PATRICIA ARIAS-CABARCOS, Paderborn University, Germany
MATIN FALLAHI, KASTEL/KIT, Germany
THILO HABRICH and KAREN SCHULZE, Universitát Mannheim, Germany
CHRISTIAN BECKER, Universitát Stuttgart, Germany
THORSTEN STRUFE, KASTEL/KIT, Germany

Brainwaves have demonstrated to be unique enough across individuals to be useful as biometrics. They also provide promising advantages over traditional means of authentication, such as resistance to external observability, revocability, and intrinsic liveness detection. However, most of the research so far has been conducted with expensive, bulky, medical-grade helmets, which offer limited applicability for everyday usage. With the aim to bring brainwave authentication and its benefits closer to real world deployment, we investigate brain biometrics with consumer devices. We conduct a comprehensive measurement experiment and user study that compare five authentication tasks on a user sample up to 10 times larger than those from previous studies, introducing three novel techniques based on cognitive semantic processing. Furthermore, we apply our analysis on high-quality open brainwave data obtained with a medical-grade headset, to assess the differences. We investigate both the performance, security, and usability of the different options and use this evidence to elicit design and research recommendations. Our results show that it is possible to achieve Equal Error Rates as low as 7.2% (a reduction between 68–72% with respect to existing approaches) based on brain responses to images with current inexpensive technology. We show that the common practice of testing authentication systems only with known attacker data is unrealistic and may lead to overly optimistic evaluations. With regard to adoption, users call for simpler devices, faster authentication, and better privacy.

CCS Concepts: • **Security and privacy** → **Biometrics**; *Usability in security and privacy;*

Additional Key Words and Phrases: Brain biometrics, user authentication, usable security, electroencephalogram (EEG)

ACM Reference format:
Patricia Arias-Cabarcos, Matin Fallahi, Thilo Habrich, Karen Schulze, Christian Becker, and Thorsten Strufe. 2023. Performance and Usability Evaluation of Brainwave Authentication Techniques with Consumer Devices. *ACM Trans. Priv. Sec.* 26, 3, Article 26 (March 2023), 36 pages.
https://doi.org/10.1145/3579356

P. Arias-Cabarcos was supported by an Alexander von Humboldt Post-Doctoral Fellowship. This work was also funded by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems" (KASTEL Security Research Labs) and Germany's Excellence Strategy (EXC 2050/1 'CeTI'; ID 390696704).

Authors' addresses: P. Arias-Cabarcos, IT Security, Department of Computer Science, Paderborn University, Fürstenallee 11, 33102 Paderborn, Germany; email: pac@mail.upb.de; M. Fallahi and T. Strufe, IT Security, Department of Computer Science, KASTEL/KIT, Am Fasanengarten 5, 76131 Karlsruhe, Germany; emails: {matin.fallahi, strufe}@kit.edu; T. Habrich and K. Schulze, Universität Mannheim, Schloss Schneckenhof-West, 68131 Mannheim, Germany; emails: t-habrich@web.de, becker-karen@outlook.com; C. Becker, Distributed Systems, Institute for Parallel and Distributed Systems, Universität Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany; email: Christian.Becker@ipvs.uni-stuttgart.de.

2471-2566/2023/03-ART26
https://doi.org/10.1145/3579356

1 INTRODUCTION

The field of **Brain Computer Interfaces (BCI)** has researched and come to solutions that allow humans to communicate with machines using their brains [86]. These technologies have been especially important in the health sector, where BCIs can for example expand the interaction capabilities of people with severe paralysis [11]. But with the development of consumer-grade **electroencephalogram (EEG)** readers [28, 34, 46, 61], new opportunities appear for using BCIs in many other realms, such as entertainment or marketing [81, 88]. Indeed, low cost headsets are already being commercialized for these purposes, and we can find app stores[1] that offer brain controlled games, relaxation trainers, and several other types of applications. In this context, and further spurred by the drawbacks of using passwords for proving online identity, research on brain biometrics has recently attracted a great deal of attention.

Brainwaves – patterns of measurable electrical impulses emitted as a result of the interaction of billions of neurons inside the human brain– present particular features that make them stand out over more traditional biometrics [32, 79]. Contrary to traits like e.g., face or gait, which can be observed from the outside and potentially misused to identify users without consent [39, 85], brain activity is not observable and thus resistant to this type of surveillance. Another noteworthy aspect is that credentials based on brainwaves can be easily revoked: our brain responses vary with the stimuli, and so in the case of having brainwaves stolen, a new credential could be generated by changing its associated stimulus [44]. Besides, given that brain activity is always present in living human beings, brainwaves can strengthen authentication with intrinsic liveness detection.

But despite the benefits of brain biometrics and the emerging democratization of EEG technology, more research is needed to make brainwave authentication applicable in real-world scenarios. Currently, the vast majority of existing work is focused on medical-grade equipment, and the scarce experiments with consumer devices involve small user samples, implement basic authentication techniques (e.g., resting), and provide limited insights on usability. Furthermore, solutions are oriented to optimize particular classification models but provide little exploration of different implementation options and their practical implications. The result is a conspicuous lack of information on how to design brainwave authentication systems for different scenarios. Motivated to fill this gap, we make two fundamental contributions to move forward[2]:

- **(1) Design, implementation, and evaluation of new authentication techniques.** We focus on techniques based on the extraction of time-locked endogenous brain responses, which are known to provide higher signal-to-noise ratio than continuous EEG recordings, the common practice in related work. Apart from techniques known in the medical-grade literature, we introduce three new tasks based on cognitive semantic processing. As a main result, we are able to achieve Equal Error Rates of 7.2%, which suppose a reduction of 68%–72% with respect to previous studies, thus demonstrating the feasibility of authentication in a moderate sized population (e.g., within an SME). Furthermore, we are the first to report a comprehensive comparison of brainwave authentication tasks, including testing six classifiers, studying one-class vs two class models, considering known and unknown attackers, analyzing the relevance of features in time and frequency, measurement channels and sample duration, considering usability, and grounded on a subject pool (N = 52) that is up to 10 times larger than the sample size in previous studies. Additionally we complement this analysis with a practical comparison of authentication performance in an open medical quality dataset, showing that, though small, there is margin for improvement (EER = 1.04%).

[1] https://store.neurosky.com/collections/apps.

[2] This article is an extension of a previous conference paper: [4].

- **(2) Usability study.** Generally, achieving high classification accuracy at the cost of low usability in authentication system design is problematic, since it can limit real-world applicability. Despite its importance, only two works so far have considered usability in the field of consumer-grade brainwave authentication. Chuang et al. [22] conducted an experimental user study asking participants (N = 15) to rate authentication tasks according to how enjoyable, easy, or engaging they were. Besides this pioneer study, Sohankar et al. [72] analyzed the usability of brainwave authentication systems in the literature against a heuristic metric built on parameters such as the type of headset or the estimated time to authenticate, but without considering users' experiences and perceptions. Here, we explore the usability of the proposed authentication techniques through empirical evidence as in [22], but extending the scope of the evaluation to: (1) cover both the usability of the tasks and the brainwave device, and (2) explore attitudes towards acceptance. Our results extend and complement previous work and aid in understanding the usability-security tradeoffs to take into account when implementing an authentication system.

Apart from these two studies, we distill lessons learned to inform future designs and research on brainwave authentication, publishing our dataset to facilitate replication and encourage further research.

The remainder of this paper is organized as follows. Section 2 introduces the status quo on brainwave authentication, defines important concepts, and sets up our application scenario and threat model. Sections 3 and 4, focus on the design of authentication tasks to collect brainwaves and detail data processing steps. We report performance and usability results in Sections 5 and 6. Finally, the paper wraps up with a discussion of lessons learned in Section 7 and conclusions in Section 8.

2 BACKGROUND

To set the background knowledge for the rest of the paper, we describe here the state of the art in brainwave authentication systems, a primer on their key components, and the threat model and use-case we adopt.

2.1 Related Work

Since the first human electroencephalogram was recorded in 1924 [33], many studies have shown that brain activity contains individuating patterns due to the influence of both genetic factors, e.g., given the unique folding structures of the cortex, and non-genetic factors, such as intelligence or previous experiences [12, 53, 86]. On these grounds, researchers have investigated the usage of brainwaves as biometrics for user identification and authentication. However, the vast majority of this research [32] has been conducted using medical-grade EEG equipment, which is highly precise, but at the same time expensive, bulky, and difficult to use. In this line of work, Palaniappan and Mandic [64], in 2007, recorded the EEGs of 102 subjects and applied classification algorithms demonstrating an overall authentication accuracy of 98%. This study and similar works have shown promising results and opened the door to further research with the advent of consumer-grade EEG devices in 2007. At this point, with low-cost, easy, and even aesthetic wearables, brainwave-based authentication for the masses has become a tangible possibility. And so the question arises whether it is possible to get accurate results with this type of EEG headset.

The literature on consumer-grade EEG authentication is scarce (see Table 1 for a structured summary), and so far it only includes experiments with a small number of subjects[3] as opposed to the

[3] Generally ≤ 10; the maximum reported number of users is 31 [2].

Table 1. Chronological Summary Brainwave Authentication Works using Consumer-grade EEG Headsets

	Data Acquisition				Data Processing			Evaluation		
Work	**Headset**	**Task**	**#Ch**	**Bands**	**Pre-processing**	**Features**	**Alg.**	**#Sbjs.**	**#S**	**Performance**
Miyamoto et al., 2009 [55]	n.a.	Resting (EC)	1	α	Spectral analysis	Spectral variance, non-dominant power spectrum	Similarity	23	1	GAR:79%
Ashby et al., 2011 [6]	Emotiv EPOC	Resting (EC), Motor + non-motor imaginary	14	$\alpha, \beta, \gamma, \delta, \theta$	Elliptic high-pass filter	AR, PSD, PS, IHPD, IHLC	one-vs.-all SVM	5	1	ACC: 100%
Nakanishi et al., 2011 [57]	n.a.	Resting, simulated driving	1	α, β	Spectral analysis	FFT, mean PS, mean PS difference between tasks	Similarity	10	10	EER: 24%
Svogor & Kisasondi, 2012 [76]	NeuroSky MindWave	Relaxation, Concentration	1	α, β	n.a.	MindWave metrics for relax and focus	Similarity	6	1	n.a.
Klonovs et al., 2013 [38]	Emotiv EPOC	Visual stimuli	4	$\alpha, \beta, \gamma, \theta$	Butterworth bandpass filter	ICA, PSD, Wavelet Analysis, zero-crossing rate	Similarity	n.a.	n.a.	n.a.
Chuang et al., 2013 [22]	NeuroSky MindSet	Resting (EC), motor/non-motor imaginary, auditive/visual stimuli	1	α, β	Extract α, β bands	PS, FT, 5-second recording windows, signal fusion, signal similarity	Similarity	15	2	HTER: 1.1%-43.3%
Mohan-chandra, 2013 [56]	Emotiv EPOC	Meditation, non-motor imaginary (math task)	14	α, β, γ	Extract α, β, γ bands	PS, PCA (only signals with >85% of signal variance), PSD, FT	Similarity	n.a.	n.a	n.a
Johnson et al., 2014 [36]	NeuroSky MindSet	Same as in [22]	1	α, β	Extract α, β bands	Same as in [22]	Similarity	18	n.a.	HTER: 1%
Nakanishi & Yoshikawa, 2015 [60]	n.a.	Route tracing, simulated car-driving	1	α, β	Spectral analysis	FFT, spectra normalization, PCA	one-vs-one SVM	30	10	EER: 22%-24%
Sohankar et al., 2015 [72]	NeuroSky MindWave	Resting	1	α	n.a.	FFT	Naïve Bayes	10	1	ACC: 95%
Chuang & Chuang, 2016 [21]	NeuroSky MindWave	Visual stimuli, mental task	1	$\alpha, \beta, \gamma, \delta, \theta$	n.a.	PS, Similarity of PS, Time windows	Similarity	10	1	FRR: 27.8%
Abo-Zahhad et al., 2016 [2]	NeuroSky MindWave	Eye blinking, resting (EC), visual stimuli	1	$\alpha, \beta, \gamma, \delta, \theta$	Elliptical band-pass filter	Eye blinking signal, AR, Visually Evoked Potentials	Discriminant Analysis	31	1	EER: 0.89%
Bashar et al., 2016 [8]	Emotiv INSIGHT	Resting (EC)	5	$\alpha, \beta, \gamma, \delta, \theta$	Band-pass FIR filter	Multiscale shape descriptor, Wavelet Packet Decomposition	Multiclass SVM	9	n.a.	TPR: 94.44%
Kavitha et al., 2017 [78]	Emotiv EPOC+	Self-related visual stimuli	14	$\alpha, \beta, \gamma, \theta$	Bandpass Filter (0.5–45Hz)	FFT, IHPD	Similarity	4	2	FAR, FRR: 12.5%
Maruoka et al., 2017 [52]	Emotiv EPOC+	Auditory stimuli (ultrasound)	2	α, β	n.a.	FFT with Hamming Window	Similarity	5	1	n.a.
Nakanishi et al., 2017 [58]	Emotiv EPOC+	Auditory stimuli (ultrasound)	14	α, β	n.a.	FFT with Hamming Window, PCA (3 best features)	one-vs-all SVM	10	10	EER: 4.4%-26.2%
Nakanishi et al., 2019 [59]	Emotiv EPOC+	Invisible visual stimuli	14	α, β, γ	ERP Extraction	PS differences for varied intensity stimuli	Similarity	20	10	EER: 23%
Zhang et al., 2020 [91]	Emotiv EPOC+	Resting EEG, gait biometrics	14	δ	Butterworth bandpass filter	Attention-based RNN	KNN	7	3	FAR:0% FRR:1%

Legend: #Ch = no. of channels, **Alg.** =Algorithm, **#Sbjs.** = no. of subjects, **#S** = number of sessions, **n.a.** = not available. Descriptions of the reported performance metrics and signal processing techniques can be found in [32, 74].

medical case. This is an important gap, since the reported accuracy may not hold when applied to larger populations where the probability of finding similar users increases [35]. Additionally, existing works mostly implement authentication based on continuous EEG recordings (e.g., while relaxing or imagining something), but few of them [2, 58, 59, 78] have looked specifically at the extraction of time-locked brain variations that appear in reaction to external stimuli. These variations, called **ERPs (Event Related Potentials)**, have been successfully tried in research with medical EEG equipment, and they are appealing for the consumer scenario given their higher signal-to-noise ratio. Another important limitation in current research is that most publications test one authentication task but there are few comparisons between different alternatives and just Chuang et al. [22] have addressed the usability of EEG authentication as perceived by users, a key aspect to understand adoption.

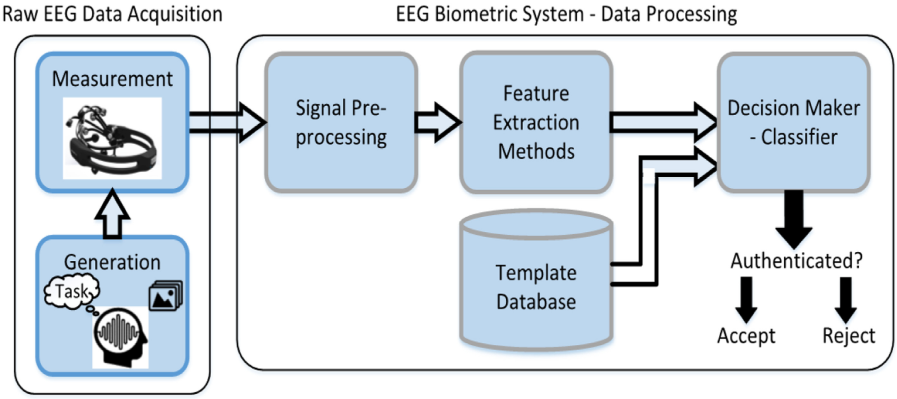


Fig. 1. Structure of a brainwave authentication system.

Looking at the existing gaps, in this work we aim to move beyond the state of the art by expanding three main fronts. First, we implement new authentication tasks based on ERPs for consumer brainwave readers. Second, we thoroughly compare these tasks, evaluating not only their performance under different attack scenarios and considering several influencing factors, but also conducting a user study to understand usability. And third, we do our experiments on a larger set of users (N = 52) and release the dataset to allow for replication and further research.

2.2 Brainwave Authentication Basics

In a biometric authentication system, users are granted access depending on their distinct physiological or behavioral traits, such as the commonly used fingerprints, voice, or face features. These traits are collected through specific sensors, processed, and compared to a previously stored sample or template from the user trying to authenticate, checking if it is a match or a mismatch. Though brainwave patterns can be used to prove a person's identity, their acquisition differs with respect to other biometrics: they need to be "generated" while performing a specific task or as a response to a stimulus, such as sounds or images. Conversely, the primary modules of a brainwave-based authentication system [32], depicted in Figure 1, are:

Generation and measurement (Section 3). Executes the *acquisition protocol* or *task* that triggers unique brainwave activity and records the associated voltage fluctuations.

Signal Pre-processing (Section 4.1). Treats the raw EEG signal to remove undesirable artifacts, such as interferences from nearby electronics, and increase the signal-to-noise ratio.

Feature Extraction (Section 4.2). Isolates the signal components that are relevant for authentication, i.e., those that contain the most information about a subject.

Classification (Section 4.3) Implements algorithms to tell authentic and non-authentic users apart.

2.3 Use Case and Threat Model

We consider a brainwave-based authentication system that protects access to applications in a desktop or laptop computer. First, the users must complete an *enrollment phase*, where their brain signals are collected to build a classification model and stored with their identity (e.g., a username). Then, during the *authentication phase*, a user supplies her identity and receives a series of visual stimuli. The generated brain responses are compared to the stored user model for denying or granting access. Therefore, for each user with true identity ID_t and claimed identity ID_c, we test the hypotheses:

$$H_0 : ID_t = ID_c \quad vs. \quad H_1 : ID_t \neq ID_c \tag{1}$$

to decide if the user is genuine or not (accept/reject H_0).

In this scenario, we consider a “zero effort” adversary [49]. This type of attacker tries to impersonate a valid user by claiming the target’s identity (ID_u) and presenting the attacker’s own biometric characteristic to the system. This adversary type can be further divided into *closed-set* and *open-set* subtypes. In the closed-set scenario, attacker samples are already available to the authentication algorithm, e.g., if the attacker is a registered user of the system. In turn, the open-set scenario considers any type of unknown attacker, whose data have not been previously seen by the authentication system [24, 87].

We assume the attacker has physical access to the device of the target victim. The resistance of a biometric system to zero-effort attacks is the system **false accept rate (FAR)**, which we calculate, among other metrics, to discuss the performance of the proposed authentication mechanisms. We use this scenario and attacker model to guide our experiments and we further discuss the applicability to different use-cases in Section 7.

3 BRAINWAVE DATA ACQUISITION

In the first step of a brainwave authentication system, specific brain signals of a user need to be activated in order to generate her credential or authentication material. This process is called acquisition protocol and can be accomplished through different types of tasks [32]. *Resting tasks*, where the user is asked to relax in a comfortable position without moving or thinking of anything in particular, are the easiest to perform. Indeed, they were among the first protocols to be investigated [68] due to their simplicity. A second category of protocols is that of *mental tasks*. In this case, users are asked to carry out imaginary actions, motor-related or not. When performing motor imaginary actions, users have to imagine kinesthetic movements of selected body parts, as opening and closing a fist or moving a finger [22]. Non-motor imaginary, on the contrary, refers to all other mental tasks that are not related to movement [90], such as mental letter composition [62], imagined speech [16], or mental calculation [56]. The last category of protocols, *stimulus-related tasks*, consists of approaches that expose subjects to stimuli of a different nature (e.g., visual, auditory, emotional).

The most common approach for brainwave authentication is to use the continuous EEG signal associated to the whole duration of a task. But stimulus-based tasks offer an alternative possibility because they can also evoke specific time-locked potentials. These brain responses, called ***Event-related Potentials* (ERPs)** [86], appear as a temporary variation of the brainwave’s voltage amplitude [40]. While more complex to implement, acquisition protocols based on ERPs provide a higher signal-to-noise ratio, being less sensitive to background perturbations [5]. This feature makes ERPs specially suitable for systems based on consumer-grade EEG devices, in which cheap sensors capture signals with lower quality compared to medical-grade electrodes [7, 27, 32]. Given the potential for ERPs to provide better accuracy, we design our tasks based on them.

3.1 Experiment Design

We focus on endogenous ERPs, a type of potential that occurs after the cognitive processing of sensory stimuli, i.e., later than 100ms after stimulus presentation.[4] While exogenous ERPs appear

[4] A comprehensive overview of currently known ERPs identified in neurological research can be found in [75].

earlier and just depend on physical parameters of the stimulus (e.g., light intensity), endogenous ERPs are partially influenced by the subject's knowledge, motivation level, and cognitive abilities [12], and so are more likely to exhibit individual characteristics useful for authentication [86]. These characteristics, together with the stable morphology of ERPs [5, 13], are the foundations for the uniqueness of these types of brainwaves. The most relevant **endogenous ERPs** are the P300 and the N400:

P300. It is a positive wave that peaks around 300ms after exposure to a certain stimulus [40]. This wave is triggered if a subject decides consciously or unconsciously that a presented stimulus or event is rare. In experimental setups, a P300 response can be elicited using the *Oddball Paradigm* [73], in which low-probability target items (e.g., pictures) are mixed with high-probability non-target or "standard" items.

N400. It is a negative wave that peaks at 400ms after a stimulus [42]. While the P300 is related to the attention of a subject, the N400 appears related to tasks that require semantic processing [40], such as language processing.

We devised five acquisition protocols to elicit the described potentials for authentication. The first two protocols focus on the P300 ERP, and were selected based on their successful application with medical-grade equipment. Besides, to further explore the space of possibilities, we introduce three new tasks built on the N400 potential that have never been used for authentication. The following list describes how we implemented the **acquisition protocols** grounded on neuroscience research techniques to trigger ERP potentials [26, 40–42, 73] :

P300:Selected. This task elicits the P300 potential based on the *oddball paradigm*. We first let the user pick a picture of her choice, which will be the target stimulus. The authentication task consists of looking at a sequence of images where the target image appears infrequently. Upon appearance, because it is a rare occurrence, a P300 is evoked that differs across subjects. To increase the attention and therefore the wave amplitude, we instructed the users to count the occurrences of the target stimuli.

P300:Assigned. This task works as the P300:Selected with the only difference that the rare picture is assigned to the user instead of being freely chosen.

N400:Words. This task is based on a *semantic priming paradigm*. Priming is defined as *"an improvement in performance in a perceptual or cognitive task, relative to an appropriate baseline, which is caused by previous, related experience"* [80]. Simply put, a subject is primed on an object if it has previous experience with this object. After priming, if the subject is presented with a semantically related stimulus, the brain finds it more meaningful and so the N400 potential appears. In our experiment, subjects watch a 'priming video' that displays cars driving on a highway. Afterward, several words are shown on the screen. A minority of these words is strongly related to the priming objects and aims at triggering N400 responses, and the rest are randomly generated.

N400:Sentences. This task is based on the concept of *incongruent sentences*. The N400 has been proved to appear when subjects read sentences word by word that end in a semantically incongruent manner [41]. An example for such a sentence is: *"Steve sat down to eat his **car**"*. Furthermore, the amplitude of the N400 wave depends on the subject's expectancy for the final word. This means that if subjects are primed on certain congruent endings, the N400 response is stronger when the incongruent word appears [42]. We therefore base on this observation to build our experiment. The task consists of showing users a sequence of sentences with slight variations. First, the sentences have semantically congruent endings, but the last variation finishes with an incongruent word to elicit a strong N400.

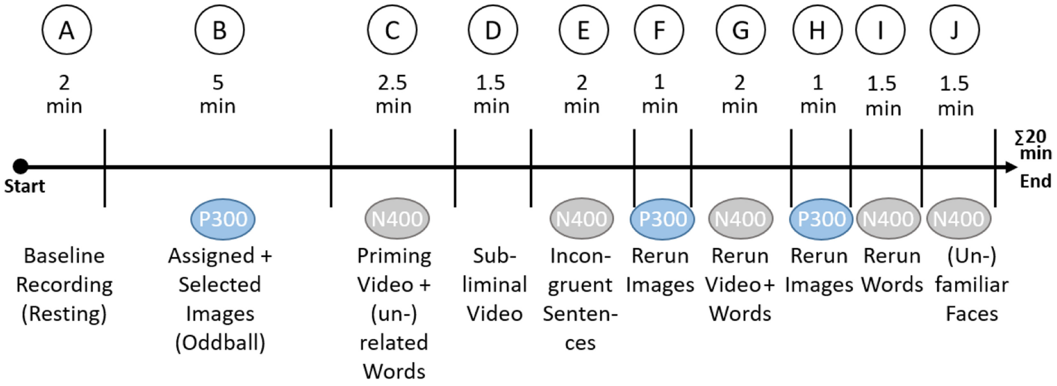


Fig. 2. Graphical flow of the experiment tasks to record users' brainwave activity for authentication. Each task is briefly described, labeled with the potential meant to be evoked (P300 or N400), and tagged with its duration.

N400:Faces. This task is based on the concept of *inhibition of knowledge* associated to N400 potentials evoked during face identification, which is another type of cognitive semantic processing, different from words. Previous work has determined that the amplitude of this wave is stronger when looking at an unfamiliar face after being presented (and therefore primed) with a sequence of familiar faces [26]. The reason is that when seeing familiar faces, the brain activates semantic representations useful to cognitively process and identify them, but these representations need to be removed and new ones activated when we start to process a new and unfamiliar individual. This inhibition of knowledge intensifies the N400. On these grounds, our protocol shows unfamiliar faces within sequences of likely familiar faces (celebrities).

3.2 Experiment Execution

Goal and Structure. The experiment at the core of this research has two goals: (1) eliciting and recording ERPs with individuating features to be used for authentication; and (2) collect information on the perceived usability of brainwave authentication.

Figure 2 illustrates the brainwave collection part of the experiment, based on the five acquisition tasks described in Section 3.1. After providing consent to take part in the study, participants were told to sit comfortably and move as little as possible during the experiment. Every room was kept rather dark and quiet, in order not to disturb the subjects. Next, their brainwave activity was recorded while performing the authentication tasks.

As shown in Figure 2, the recording starts with baseline measurements of brain activity while resting. Then, it follows with several sequences and repetitions of the authentication tasks[5], to acquire multiple samples for training and testing the classification algorithms. After the recording, participants filled out a paper questionnaire to assess the usability of a brainwave authentication system based on the performed tasks and headset (details in Section 6). All experiment materials are linked in Appendix A.

Apparatus. We use the Emotiv EPOC+ headset [28] to record brainwave activity. We chose this device because it is the prevalent choice in scientific studies and it offers a higher number of recording channels (14) than other consumer grade products, which leads to more accurate mea-

[5] Element D in the study flow depicted in Figure 2 was included to test subliminal manipulations. Since we did not obtain conclusive results in this regard, we just report it as a study item without giving further details.

surements.[6] The experiment flow was programmed with PsychoPy [66], an open source tool for conducting experiments in behavioral sciences, and connected to the EPOC's reading software to synchronize stimuli presentation with brainwave recording. The recorded data was processed using the open Python libraries `MNE`[7], for analyzing EEG signals, and `scikit-learn`[8], for programming the classification algorithms.

Recruitment and Ethical Aspects. We recruited participants following a self-selection sampling approach [43]. The study was advertised through different channels asking for volunteers, including online posts, flyers spread at different university locations and brief announcements during lectures. Each participant received information about the experiment and about how we would treat their personal data fulfilling the EU **General Data Protection Regulation (GDPR)** [29], in order to get informed consent. To avoid biasing the subjects, we disclosed the actual purpose of the experiment, i.e., building an authentication system, at the end of the recording session and before the usability questionnaire. The approximate average duration of the whole study was 45 minutes and we compensated participants with 5€ and a report on their brainwaves containing information about interest, stress, and focus level during the study. Subjects were also told that participation was voluntary and the experiment could be abandoned at any time. The whole procedure is IRB-approved.

Participant Demographics. In total, 56 subjects took part in the experiment, conducted between May 8 and July 2, 2019. We recorded ERPs from 23 females (41.1%) and 33 males (58.9%), leading to a slightly imbalanced gender distribution. With regard to age, our population is skewed towards young adults because most of the experiments were conducted with university students. The majority, 28 subjects (50%), fall in the age range 18-24, followed by 16 (28.6%) participants aged between 25 and 31, and 8 (14.6%) in the range 32-38. The remaining 4 persons (7.2%) were over 39 years old.

4 BRAINWAVE DATA PROCESSING

To have useful brainwave data for the classification algorithms that implement authentication, raw EEG signals undergo a two-step preparation process to: (1) remove undesirable artifacts, and (2) extract relevant features for authentication. This section summarizes the data preparation steps we followed, based on common practice in the literature [32], and the classification models we apply to these data.

4.1 Pre-processing

We require pre-processing brainwaves to remove noise, reject external interference such as blinking and achieve a fixed vector for the feature extraction step. The data collected during the experiment consists of continuous EEG recordings lasting approximately 20 minutes and sampled at a rate of 256Hz. However, authentication is required for only specific relevant sections surrounding the presentation of stimuli, i.e., the ERP waves. These sections, termed *epochs*, comprise a user sample.

First, we applied a *Finite Impulse Response bandpass filter* of 1–50Hz [63] to eliminate the electrical noise generated by 50Hz power lines and concentrate on the most useful brain activity, which is primarily contained within this frequency range [32]. Following that, to extract the ERPs, we cut 1-second epochs from 100ms before stimulus presentation to 900ms afterward to ensure that we

[6] The reader is referred to [71] for a comprehensive review and comparison of consumer grade EEG readers, including research applications.

[7] https://mne.tools/stable/index.html.

[8] https://scikit-learn.org/.

obtain all of the potential's information, taking into account variances in peak latency [75]. Meanwhile, to make the data comparable for classification and to facilitate feature extraction [38], the baseline for each epoch must be removed. This is accomplished by subtracting the sample's moving average from 100ms before the stimulus's presentation. Baseline removal is performed for each epoch and channel separately. The resulting signal is distributed around 0μV at the start of stimulus presentation. Finally, we eliminated epochs with large artifacts contaminating the EEG signal, such as voltage changes caused by eye or muscle movements. Typically, thresholds of around 100μV are used to remove these artifacts [32]. Since the EPOC+ provides lower signal quality with increased noise than typical medical-grade devices [6], we set the voltage threshold to 150μV in our case.

4.2 Feature Extraction

The next step, using the clean EEG signal, is to obtain discriminant features that represent and encode the user's mental activity [32]. We selected the most frequently used features in the *time* and *frequency* domains applied in previous work [1, 3, 32, 89], and used them as a starting point for determining which features are most appropriate for our proposed tasks (Section 5.2.5).

First, we fit the ERP epoch, which is a 1-second time-series, to an ***Autoregressive*** **(AR)** model and consider their coefficients as features. In a time series, the most recent values are influenced by the previous values to some extent [32]. As a result, the *Autoregressive* (AR) model is commonly used in EEG research [89]. The AR model is a time-domain representation of a random process in which the output variable is linearly dependent on its own previous values and a stochastic term that is not perfectly predictable. For a random process X_t, the AR model of order p can be defined as follows [3, 32, 89], where a_i denotes the AR coefficient of the model at delay i, X_t denotes the current value of one channel, and ϵ_n represents the white noise at time n (see Equation 2). AR coefficients have the potential to reveal certain intrinsic properties of the EEG signal within a single channel, making them an attractive candidate for extracting subject-specific information from recorded signals.

$$X_t = -\sum_{i=1}^{p} a_i X_{t-i} + \epsilon_n \tag{2}$$

Second, we determine each epoch's **Power Spectrum (PS)** across multiple frequency bands (low [1–10Hz], α [10–13Hz], β [13–30Hz], and γ [30–50Hz]), by employing the Welchs periodogram algorithm to calculate the **Discrete Fourier Transformation (DFT)**. Ashby et al. used a similar strategy to segment their EEG recordings into individual epochs. As a result, each channel has four features, and each epoch has 56 features as we recorded EEGs on 14 channels. An example input and output of the discrete FT is depicted in Figure 3.

4.3 Classification

For the purpose of authentication, the recorded data samples of each user need to be compared to stored samples of the same subject and classified as matching or not. In normal identification settings, every sample shown to the system will be assigned to a class. Such a system cannot be used to perform authentication because some samples need to be excluded by the system. As a result, a different method for implementing an authentication system must be selected. The basic idea is to learn based on a *one-vs.-all classification* approach, using two classes. Accordingly, a classifier is trained for each subject that will be included in the system. As a result, a single classifier is responsible for recognizing a single subject. We compare and discuss the applicability of these authentication approaches for different classifiers under two different attack scenarios.

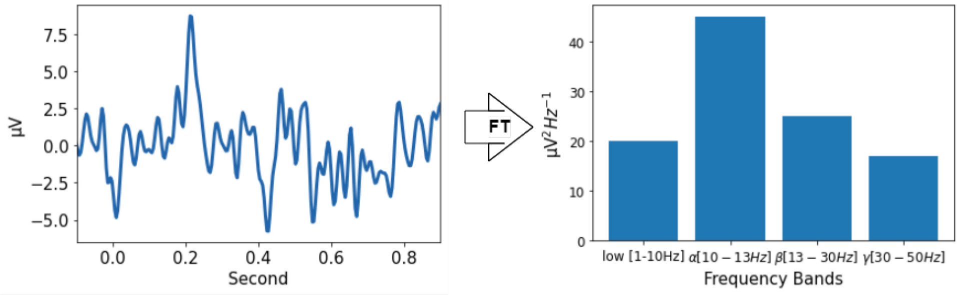


Fig. 3. In- and output of the transformation from time domain (left graph) into frequency domain (right graph) using the discrete Fourier Transformation (FT) for bands low, α, β and γ.

Table 2. Brainwave Datasets for Five Authentication Tasks

Dataset	**#users**	**#samples**
P300:Selected	49	819
P300:Assigned	49	803
N400:Words	49	1484
N400:Sentences	44	238
N400:Faces	46	399

4.3.1 Classifiers. Various Machine Learning techniques have been used in the literature, ranging from simple *threshold methods* with low computational effort [55], to more powerful methods such as ***Linear Discriminant Analysis*** **(LDA)** or ***Support Vector Machines*** **(SVM)** [45]. According to Lotte et al. [45], *Deep Learning* techniques are not yet fully applicable in the EEG authentication context due to the typically small size of brainwave datasets. We tested our feature set with a set of representative algorithms suited to our dataset size, in order to establish a baseline: LDA, SVM, **Gaussian Naive Bayes (GNB)**, **k-nearest neighbors vote (KNN)**, **Random forest (RF)**, and **logistic regression (LG), Dummy classifier (DC)**.[9]

Regarding the classification model, despite one-class classifiers offer the advantage that only require training data from the genuine user, our previous work [4] showed they cannot provide minimally acceptable results, operating close to a random classifier. As a result, we decided to disregard this model and focus only on two-class classifiers, which have potential to bring us one step closer to practical brainwave authentication with consumer-grade devices.

We applied the selected classifiers for user authentication with the five authentication tasks defined in Section 3. To increase the reliability of the classification results, we used repeated stratified k-fold cross-validation with (k = 3). For this reason, we removed subjects with fewer than three epochs per task from the dataset to have at least two samples for learning and one for testing in each round of cross-validation. The final datasets are listed in Table 2 and linked in Appendix A. The number of repetitions was 10 and we report the mean result across all folds from all runs. We scale our features and, in order to avoid overfitting, StandardScaler[10] was fitted on the training set and then applied to the train and test sets in each round.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.
[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

4.3.2 Attack Scenarios. The authentication system was implemented and evaluated under two scenarios: *closed-set* vs *open-set*. For the *closed-set* scenario, we followed a standard *one-vs.-all* approach. According to this scheme, we built specialized classifiers for each user by labeling all its samples as "authenticated", and assigning the "rejected" label to all samples coming from other users.

While this is the scheme implemented by the majority of papers, the model could learn information regarding rejected users during the training phase. However, the assumption that the attacker data has been seen by the classifier is unrealistic, as authentication systems generally do not have access to attacker epochs in the real world. To develop a more practical authentication system, we evaluate it under the *open-set* scenario. To test this setting, we divided our dataset samples into "authenticated", "rejected", and "attacker" in each round of cross-validation, grouped based on the subject ID. Authenticated subjects' epochs were used in both training and test sets, and rejected epochs were used for the training set, while the attacker samples were only used in the test set. For example, suppose we have 39 subjects. In that case, a user's epochs get authenticated labels, 29 rejected, and 9 attackers per model (train test splits 75% and 25%, respectively.) It is worth mentioning that we have multiple samples/epochs per user in the dataset. The main idea of the *open-set* scenario came from Buschek et al.'s evaluation of keystroke biometrics [18], in which 4.7–25.1% lower EERs are achieved when attacker samples are included in the training set, illustrating the relevance of considering realistic attack scenarios.

5 AUTHENTICATION PERFORMANCE

In this section, we evaluate the performance of the proposed authentication tasks and classification algorithms, analyzing feature relevance and other factors influencing performance. We contextualize the results theoretically regarding performance values reported in related work; and also practically, by implementing and evaluating our system on an open-access medical-quality dataset.

5.1 Evaluation Metrics

Several methods can be applied to evaluate classification systems. In the case of a binary problem, there are four possible classification results: (1) authenticate a legitimate user **(True Positive, TP)**, (2) authenticate an illegitimate user **(False Positive, FP)**, (3) deny an illegitimate user **(True Negative, TN)**, and (4) deny a legitimate user **(False Negative, FN)**. Based on the frequency counts of these results, the performance of the system is typically assessed by its **False Acceptance Rate (FAR)**, **False Rejection Rate (FRR)**, and **Accuracy (ACC)**. The FAR compares the number of false positives to the sum of false positives and true negatives, i.e., how often an impostor is authenticated as legitimate. In turn, the FRR compares the number of false negatives to the sum of true positives and false negatives, giving an idea of the frequency at which the system rejects legitimate users. Finally, the ACC represents the number of correct predictions over the total number of predictions made by the classifier.

The selection of metrics and how to report them is crucial to properly evaluate a system. Thus, since the accuracy is not meaningful to evaluate highly imbalanced classification problems like the authentication scenario, where we have one category representing the overwhelming majority of the data points, we focus on the metric pair FAR-FRR. These metrics, however, are tied to a specific configuration of the classification threshold. To understand the full operational range of the classifiers, we visualise results with **Receiver-Operating-Characteristic (ROC)** curves, which plot the FAR and True Positive Rate, TPR (=1-FRR) as a parametric function of the threshold. We also report **Equal Error Rates (EER)**, as a summary metric that represents the point where FAR and FRR are equal. This reporting scheme, as suggested by Sugrim et al. [74], allows for a better under-

Table 3. Average Equal Error Rate (EER) for 5 Authentication Tasks, Comparing Closed-set vs Open-set Attacker Scenarios

					Equal Error Rate (%)			
					Classifiers			
Task	Scenarios	LG	RF	KNN	GNB	SVM	LDA	DC
P300:Selected	closed-set	8.1±5.7	9±5.8	17.7±8.8	16.4±7.5	**7.6±4.6**	12.1±7.0	**50±0**
P300:Selected	open-set	10.5±6.2	10.2±6.3	18.5±9.0	16.3±7.4	**8.6±5.1**	14.4±7.5	
P300:Assigned	closed-set	7.7±4.9	9.2±6.1	18.1±10.0	16±8.1	**7.2±4.8**	10.5±6.7	
P300:Assigned	open-set	9.4±5.4	10.1±6.3	19±10.0	16.3±8.0	**8.5±5.5**	13.1±7.3	
N400:Words	closed-set	8.8±6.5	8.9±6.5	17.5±9.6	15.5±6.8	**8.4±5.8**	10±7.0	
N400:Words	open-set	11.4±7.1	10±6.8	18.3±9.6	15.5±6.4	**9.4±6.1**	13.2±7.7	
N400:Sentences	closed-set	**11.5±8.0**	13.7±8.7	21.6±12.5	25.2±10.6	11.7±11.5	17.0±11.9	
N400:Sentences	open-set	13±8.6	15.2±10.1	22.2±12.6	26.3±10.2	**12.3±11.6**	19.7±11.5	
N400:Faces	closed-set	9.8±8.0	10.8±6.8	16.6±9.6	18.9±10.0	**8.9±6.6**	14.7±10.5	
N400:Faces	open-set	13.9±8.3	11.9±7.3	17.6±9.9	18.2±10.8	**9.8±7.5**	17.5±10.7	

Best classification results for each task/scenario are highlighted in bold, and the overall best result, with a green background.

Table 4. Average **False Rejection Rate (FRR)** at 1% **False Acceptance Rate (FAR)** for Five Authentication Tasks, Comparing Different Classifiers and Attacker Scenarios

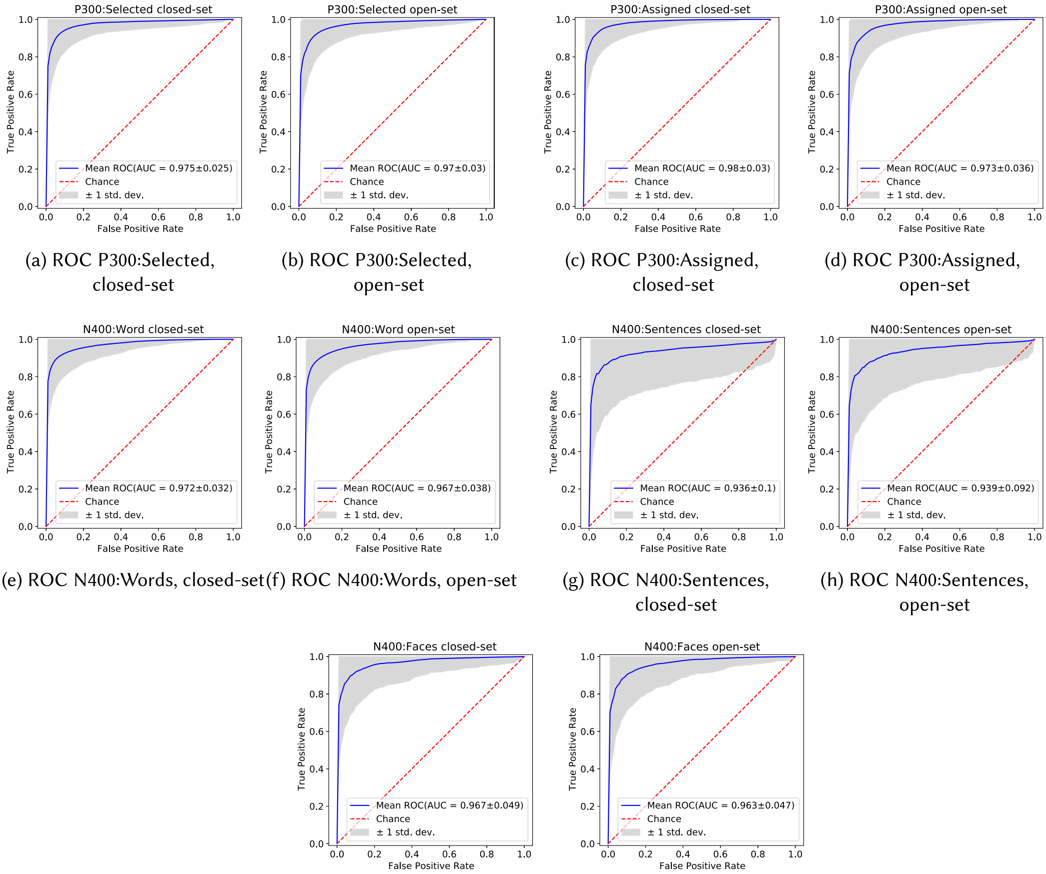
					FRR at FAR = 1% (%)			
					Classifiers			
Task	Scenarios	LG	RF	knn	GNB	SVM	LDA	DC
P300:Selected	closed-set	33.1	28	42.5	58.3	**25.3**	53.9	**99**
P300:Selected	open-set	49.4	32.6	49.8	58.4	**29.3**	65.6	
P300:Assigned	closed-set	32	29.5	42.8	52.8	**24.3**	48.8	
P300:Assigned	open-set	49.6	34.2	49.9	52.9	**29.2**	63.1	
N400:Words	closed-set	30.8	26.2	37.1	51.3	**22.4**	44.5	
N400:Words	open-set	51.9	29.4	46.5	54.9	**26.8**	59.9	
N400:Sentences	closed-set	49.8	47.3	51.1	64.1	**34.4**	64.4	
N400:Sentences	open-set	59	51.2	55.3	65	**38.5**	73	
N400:Faces	closed-set	37.9	33.9	43.3	50.3	**25.7**	51.1	
N400:Faces	open-set	55.1	37.5	49.1	51.4	**28.8**	64.7	

Best classification results for each task and scenario are highlighted in bold. The overall best result is signaled with a green background.

standing of the operation capabilities of authentication methods, and how they can be configured for different use-cases. While ideally both FAR and FRR should be as close to zero as possible [49], in real systems each metric increases at the expense of decreasing the other. A low FAR is linked to a better security level, while the FRR relates to usability. We report FRR at FAR equal to 1%, which is the minimum acceptable FAR concerning application security [19]. In addition to ROC and EER, this metric could provide a better understanding of usable EEG-based authentication systems.

5.2 Results

The results for all classifiers and all tasks are summarized in Tables 3 and 4. The best authentication results are obtained with the SVM classifier, which yields EERs ranging from 7.2% to 12.3%, and FRRs at 1% FAR in the range 22.4% to 38.4%. The only exception occurs for the N400:Sentences under the closed-set attacker scenario, for which LG performs slightly better when looking at the



(a) ROC P300:Selected, closed-set
(b) ROC P300:Selected, open-set
(c) ROC P300:Assigned, closed-set
(d) ROC P300:Assigned, open-set
(e) ROC N400:Words, closed-set
(f) ROC N400:Words, open-set
(g) ROC N400:Sentences, closed-set
(h) ROC N400:Sentences, open-set
(i) ROC N400:Faces, closed-set
(j) ROC N400:Faces, open-set

Fig. 4. Performance comparison of five authentication tasks using **Support Vector Machine (SVM)** classification. ROC curves are depicted for each authentication task in both closed-set and open-set attacker scenarios.

EER metric. However, considering the FRR at 1% FAR, it becomes obvious that SVM is a more reliable option for an authentication system, since the performance is consistently better for all tasks and scenarios. Moreover, RF could be considered the second-best classifier for our goal. These machine learning classification results are consistent with those obtained by Fernandez-Delgado et al. [30], who tested 179 classifiers on the 121 datasets from the UCI database[11], reporting that RF and SVM outperform all other options in a variety of classification problems. In the rest of the paper, we center the comparison of tasks and scenarios, as well as their applicability, on our best results, obtained with the SVM classifiers. Feature selection, as well as the effect of epoch duration and measurement channels are discussed at the end of this section.

5.2.1 Comparison between Authentication Tasks. With regard to authentication tasks, our results establish the P300 protocols as better authentication options than the N400 protocols, and the best performing task was P300:Assigned with an average EER of 7.2%. Figure 4 provides the

[11] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/index.php.

ROC curves for the best classifiers, illustrating the operational range of the five authentication models in both closed-set and open-set scenarios. Classifiers are configured to use five features per channel (one in the time domain and four in the frequency domain), as it will be justified later. The **area under the curve (AUC)** represents the probability that a random illegitimate user is scored lower than a random genuine user, i.e., how well the classifier can separate users. While the P300:Assigned outperforms the rest of the tasks in the tested conditions based on the AUC, all schemes show potential for discerning users, and therefore, could be feasible for brainwave-based authentication (AUC for the dummy classifier is 50%). However, there is variability in the average ROC curves. In this regard, an important factor to consider in the comparison is the different number of samples and users per task. As it can be observed in Figure 5(a), the N400:Words and P300 tasks have the largest number of samples (1484, 819, and 803 for 49 users). Nevertheless, for the remaining N400 protocols, the datasets are reduced to 44 users and 238 samples for the N400:Sentences, and 46 users and 399 samples for the N400:Faces. Accordingly, it can be observed that protocols with more user samples perform better, which can be attributed to a larger volume of data available for the machine learning model. For example, the performance of N400:Sentences and N400:Faces could have been negatively impacted by the small number of samples per user (5 and 8 on average), which leads to very few data for training and testing set in the cross-validation process. The results in Table 4 are consistent with this observation, showing that the best FRRs at 1% FAR are obtained for the protocol with more samples per user (30 on average), the N400:Words.

5.2.2 *Comparison between Closed-set and Open-set Scenarios.*

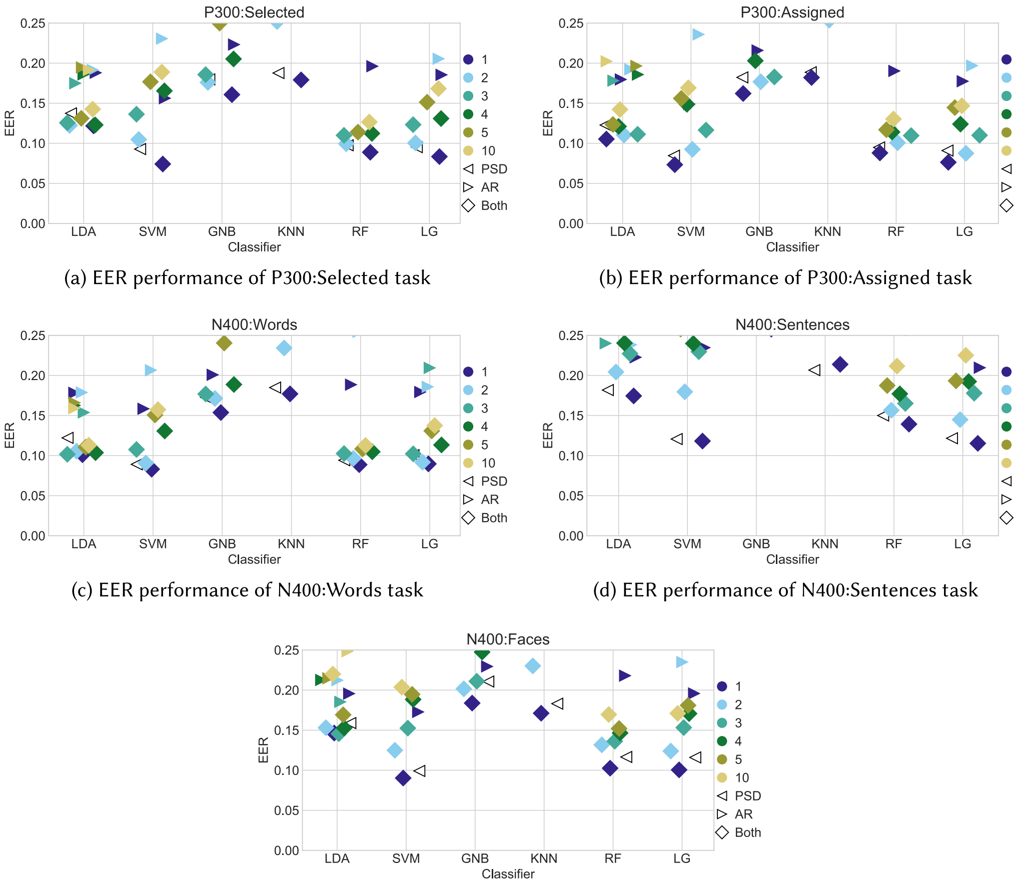
As expected, the results indicate a significant performance degradation in open-set scenarios compared to the closed-set setting. For example, as shown in Tables 3 and 4, the EER increases between 5.1–18% and the FRR at 1% FAR raises between 11.8–20%.

In this evaluation, however, the comparison might be affected by the different sizes of the attacker spaces in each scenario. Having fewer attacks in the open-set scenario might lead to overfitting threshold that would not work with a higher number of attackers. As an additional test to account for this limitation, we split the dataset into two datasets without overlapping subjects (D1, D2). In the closed-set scenario, we train/test on each dataset separately, then average the results. For the open-set scenario, we train classifiers using D1 samples and use D2 samples as attacker samples for the testing set and vice versa (train D2, test D1). The results of this test are similar to those reported in Table 3, but yield a bigger difference between the open-set and closed-set scenarios. The gap was about 50% higher for SVM as the leader classifier. Full results are linked in Appendix A.

This consistent trend of worst performance for open-sets, underscores the need to investigate this more realistic scenario when evaluating brainwave systems (not common in related work), since attacks can likely come from persons who have not pre-registered in the system.

5.2.3 *Applicability.*

In real-world authentication scenarios, systems operate not at the EER but at configuration points where the FAR is lower than the FRR to minimize the probability of impostors accessing the system. In general, most biometric systems have an FRR ranging from being falsely rejected one out of five times up to one out of a thousand times (i.e., 20% to 0.1%). The FAR is more critical for security and usually ranges from 1% for low-security applications to 0.00001% for very high-security applications [19]. Our results (see Table 4) show that the best configuration is obtained for authentication with the N400:Words task at FAR = 1% and FRR = 22% (closed-set), and FAR = 1% and FAR = 26.8% (open-set). While the FAR value is equal to the needs of low-security application scenarios, the FRR still needs improvement. More precisely, this model falsely rejects a user about one out of four times, which already shows a significant improvement compared to our previous work [4].



(a) EER performance of P300:Selected task

(b) EER performance of P300:Assigned task

(c) EER performance of N400:Words task

(d) EER performance of N400:Sentences task

(e) EER performance of N400:Faces task

Fig. 5. Performance comparison of five authentication tasks using different classifiers and features based on the **Equal Error Rate (EER)** metric. We neglected results with an EER above 25% to concentrate on the most promising results.

According to the ROC and EER standard deviation values, we observe a relatively high variability, which poses a challenge for real-world implementation. We believe that in future research, it would be worthwhile to investigate each subject individually in order to discover the reason for this variability and potential mitigation strategies. Another important consideration is that we have trained and tested the classifiers on imbalanced datasets, with more attacker than legitimate samples, which can bias the classification towards better recognizing the attacker class. To test how the classification would perform in a balanced dataset, we repeated the evaluation by oversampling the legitimate user class to have the same amount of data for legitimate users as we have for attackers, obtaining very similar performance results. In a practical setting, this would mean that there is a need to collect more user samples.

We expect lower error rates in real implementations with personalized stimuli. The reason is that we measure and report the FAR by directly comparing impostors' ERP samples to the legitimate user model. But if we consider the dynamics of the authentication protocols, those ERPs should appear in response to the target stimuli (e.g., unfamiliar faces within a series of familiar ones).

Checking this condition before accepting an ERP will yield lower FARs, as it is highly unlikely that an impostor reacts to the stimuli designed for the legitimate user. Therefore, the obtained FAR is to be understood as a rough upper bound. To further improve performance, and therefore applicability to high security scenarios, the proposed tasks could be combined in a multi-modal authentication setting or used as a second factor [2, 6, 91].

5.2.4 Effect of Feature Extraction on Performance. The process of extracting features from epochs is critical in developing an EEG-based authentication system. Based on previous work [32], we used the **auto-regressive (AR)** model and the **power spectral density (PSD)** in this step as a starting point, and tested which components have a more significant impact on the final results. We extracted combinations of features and measured their classification performance, including AR coefficients (order=1,2,3,4,5,10) of the epoch, PSD of the epoch, and a combination of the two, totaling 13 different types of feature sets per task (6 AR, 1 PSD, 6 combinations of both AR and PSD.) We focused on combinations with low number of features because a high number of features would be difficult to learn by the simple models we use and with the limited number of samples available.

The results, plotted in Figure 5, demonstrate consistent performance for the same classifier and feature set combination across all tasks. AR coefficients with a lower order yield lower EERs. While PSD is always superior to AR features, the combination of order one AR features and PSD is the optimal choice for our authentication system. Furthermore, SVM outperforms all the other models in each of the five tasks. Therefore, we used this configuration to build our authentication system.

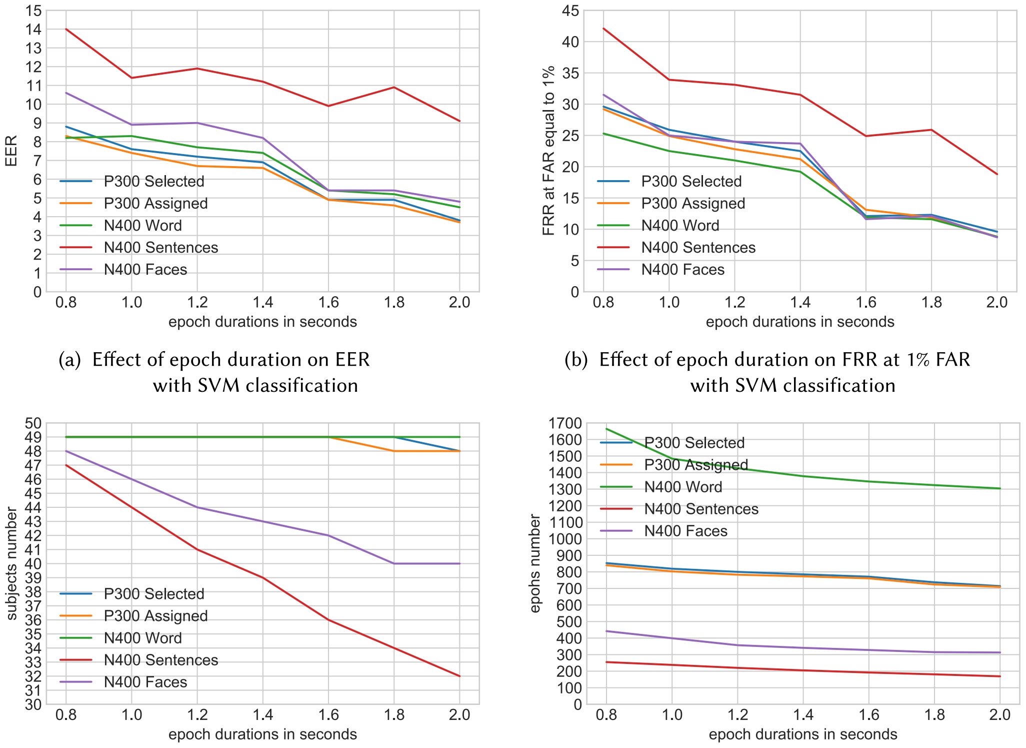
5.2.5 Effect of Epoch Duration on Performance. We evaluated our system using SVM classification and the same processing pipeline for different epoch durations to analyse the impact on performance. We configured epochs to last 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0 seconds, beginning 0.1 seconds prior to the ERP event. As illustrated in Figure 6, the duration of epochs can affect results from a variety of perspectives. Due to the similarity between the outcomes for the open and closed-set attacker scenarios, we just plot and discuss the results of the latter scenario in this section.

First, Figures 6(c) and 6(d) demonstrate that increasing the duration of epochs increases the number of rejected epochs, which may result in a decreased number of available subjects for the authentication system. This is due to the fact that, as previously stated, we eliminate the subject if they have fewer than three epochs. With longer epochs, the reason for high rejection is that it is more likely to have data points that pass the rejection threshold than in shorter ones. For example, in the N400:Sentence task, the number of epochs decreased from 255 using 0.8-second epochs to 169, using 2-second epochs, while the number of available subjects decreased from 47 to 32.

Second, we can see from Figures 6(a) and 6(b) that using epochs with a longer duration yield significantly better performance considering both the EER and the FRR at %1 FAR for all five authentication tasks. This improvement is stronger for the second metric, a crucial indicator for the usability of authentication systems. For instance, authentication with the P300:Selected task improved by reducing more than 56% in EER and more than 67% in FRR at %1 FAR (i.e., from 8.8% to 3.8%, and from 29.6% to 9.6%, respectively).

Finally, we observe a positive effect of increased epoch duration on performance, which is worthy of further investigation in future work.

5.2.6 Channel Influence Analysis. We investigated the influence of EEG measurement channels on performance by calculating the EER for each single channel. As it can be observed in Figure 7(a), all channels have similar performance for each authentication task, especially in tasks with a higher number of samples (both NP300, and N400:Words), where the variance is smaller.



(a) Effect of epoch duration on EER with SVM classification

(b) Effect of epoch duration on FRR at 1% FAR with SVM classification

(c) No. of available subjects with respect to epoch duration with SVM classification

(d) No. of available epochs with respect to epoch duration with SVM classification

Fig. 6. Effect of epoch duration on classification results and epoch rejection for 5 brainwave-based authentication tasks. Figures (a) and (b) show performance variability. Figures (c) and (d) show the changes in number of available subjects and epochs.

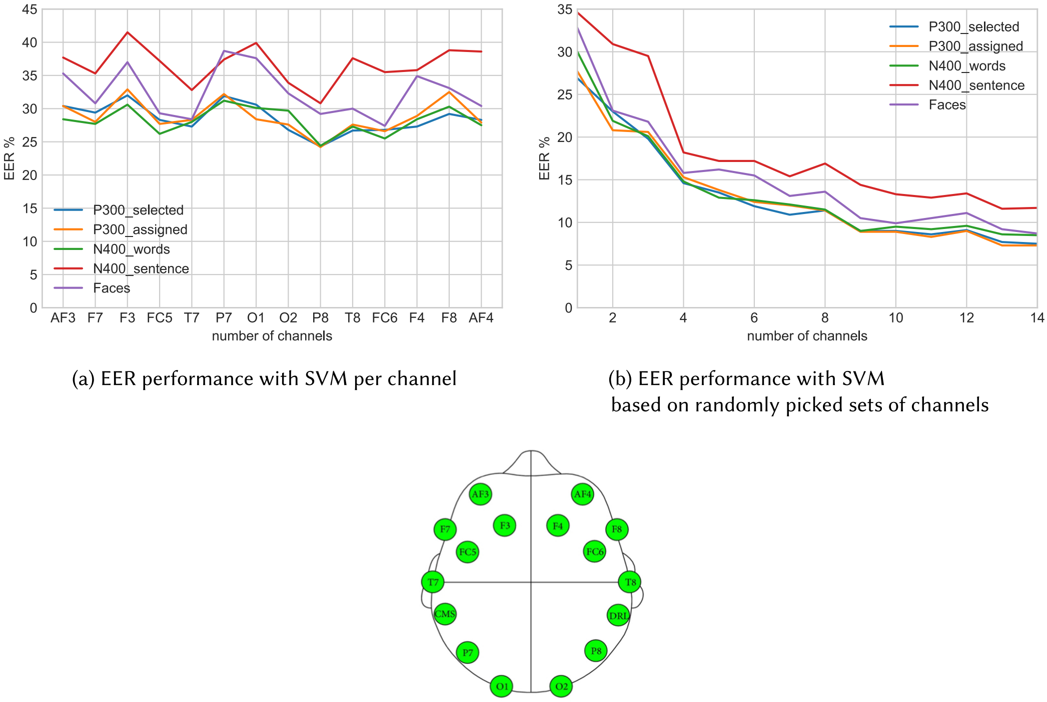
This suggests that electrode locations do not have an important effect in our authentication scenarios with consumer grade devices.

To further understand the impact of the number of electrodes in performance, we randomly picked channel combinations, varying from n = 1 to n = 14 channels, and built the authentication system based on their measurements. The effect of channel number on the EER is plotted in Figure 7(b). The results show that the average EER of the five tasks decreases from 30.4% using one channel to 8.7%, using 14 channels, i.e., a 71.3% improvement. It is apparent that, generally, more channels lead to better results; however, after n = 9 channels, performance seems to stabilize.

5.3 Contextualization of the Performance Results

In this section, we first position our contributions with regard to similar related work, as a direct comparison is not possible due to unavailability of datasets captured with consumer-grade devices. Then, we practically compare our performance results with those that would be obtained with high-quality data collected with a medical-grade device.

5.3.1 Positioning within Related Work using Consumer Devices. Comparison with existing works on brainwave authentication is challenging due to the frequent under-reporting of metrics (usually presented for an optimized configuration without providing ROCs) and the differences in the number and diversity of samples, algorithms, experimental conditions, and other aspects that



(a) EER performance with SVM per channel

(b) EER performance with SVM based on randomly picked sets of channels

(c) Channel names and location for the Emotiv EPOC+ headset

Fig. 7. Channel Influence Analysis Figure (a) each channel and Figure (b) number of channels.

influence performance. Furthermore, none of the related works in Table 1 provides an open dataset to replicate the results. Acknowledging these difficulties, we describe our results along with other relevant works in the literature that also focus on ERP-based tasks and report EER values.

Nakanishi et al. investigated various authentication tasks [55, 57–60], including low intensity visual stimuli (EER = 23%, n = 20), and ultrasound stimulation (EER = 26.2%, n = 10). In both cases, our P300:Assigned protocol has better performance under our experimental conditions. We decrease the EER from 23%–26.2% to 7.2% for the P300:Assigned task, which means a relative error reduction of 68–72%. Also, for the rest of the tasks (P300:Selected and all N400), we observe an average improvement of 68%, 65%, 51%, and 63%, respectively. These results indicate that standard visual tasks are potentially more suitable for brainwave authentication than current ERP-based proposals in the literature. For an accurate comparison, however, this should be tested under the same conditions.

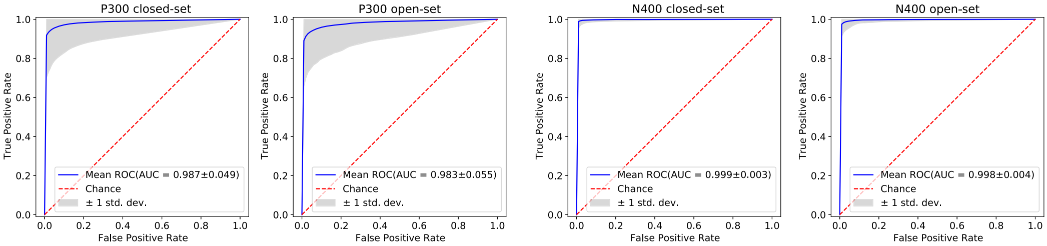
In general, works reporting the lowest EERs or FAR/FRR use multi-modal fusion or a second authentication factor [2, 6, 58, 91] (EER = 0%,EER = 4.4%, EER 2.9%, EER = 0.89%, and FAR = 0%/FRR = 1%) to complement brainwaves, which suggests these are viable paths to further improve the applicability of our tasks.

5.3.2 *Practical comparison using data collected with medical-grade devices.* To provide further insights on the performance of brainwave authentication with consumer-grade devices vs. using medical-grade headsets, we decided to evaluate our classifiers on an open high-quality EEG dataset. While we could not find datasets collected by researchers building authentication systems, there

Table 5. Average Performance of 2 Authentication Tasks on the ERP CORE Dataset [37], Comparing Closed-set vs Open-set Attacker Scenarios

			Performance (%)						
						Classifiers			
Metric	Task	Scenarios	LG	RF	KNN	GNB	SVM	LDA	DC
EER	P300	closed-set	2.6±1.9	**1.9±2.4**	6.8±6.9	8.9±4.7	3.5±3.5	4.3±3.4	**50±0**
	P300	open-set	6.7±3.0	**3±2.2**	8.9±7.1	9.5±4.4	5.2±4.1	9.8±5.1	
	N400	closed-set	4±2.7	**1.04±0.9**	20.9±9.1	10.7±3.6	6±4.0	3.4±2.7	
	N400	open-set	6.8±3.5	**1.9±1.2**	21±8.8	11.6±3.7	7.5±4.1	7.4±3.8	
FRR at FAR= 1%	P300	closed-set	5.9	**2.6**	13.3	58.4	6.5	17.3	**99±0**
	P300	open-set	34	**5.1**	26	54.5	11.3	48.2	
	N400	closed-set	11.1	**1.1**	51.5	77.8	16.9	12.2	
	N400	open-set	35.5	**2.7**	57.7	69.6	23.3	43.6	

Best classification results for each task/scenario are highlighted in bold, and the overall best result, with a green background.



(a) ROC P300, closed-set (b) ROC P300,open-set (c) ROC N400, closed-set (d) ROC N400, open-set

Fig. 8. Performance comparison of two authentication tasks using **Random Forest(RF)** on the ERP CORE dataset [37]. ROC curves are depicted for each authentication task in open-set and closed-set attacker scenarios.

are several open datasets collected for other classification tasks based on ERP elicitation. From the limited available options, the ERP **CORE** dataset (**Compendium of Open Resources and Experiments**) [37] is the best option regarding number of users and types ERPs. This dataset contains data for 40 subjects and seven widely used ERP components, recorded with the high resolution EEG reader Biosemi ActiveTwo[12] (30 electrodes) at 1024Hz frequency.

We developed an authentication system based on the ERP CORE dataset to have a baseline for comparison, using the P300 and N400 ERP components collected following an oddball paradigm protocol, and a word association protocol, respectively, similar to our tasks. We performed preprocessing, feature extraction, and classification exactly in the same way as in our testbed.

The results obtained on the ERP CORE dataset establish the N400 protocols as slightly better authentication options than the P300 protocols, and the best performing classifier was the RF. Meanwhile, SVM also showed an acceptable performance, which is again consistent with our expectations based on previous machine learning research [30]. More detailed results are presented in Table 5 and the ROC curves in Figure 8.

The better results obtained for N400 ERPs could be attributed to the more significant number of available samples for this task (2,268 samples) compared to 1,342 samples for the P300 task, as we have seen in our analysis with the consumer EEG reader. In addition, Figure 8 shows a lower

[12] https://www.biosemi.com/products.htm.

variance in all tasks, and specially in the N400 case, which is also consistent with the analysis in Section 5.2.

According to the data in Table 5, the results of unknown attacker scenarios (open-set) are worse than those for the open-set scenario. In fact, once the result improves due to the high-quality dataset, the effect of testing the open-set attack scenario becomes more apparent. For example, in the case of RF, which is our best classifier on this dataset, the open-set EER increases by 82% (from 1.04% to 1.9%), and the FRR at 1% FAR raises a 145% (from 1.1% to 2.7%). Thus, we strongly encourage researchers to apply this scenario evaluation when designing EEG-based authentication system, as it is more representative of real-world use cases and appears to significantly impact the results.

By comparing performance of consumer-grade device and medical device, we found that the best results were 22.4% and 1.1% FRR at FAR, equal to 1%, respectively. Despite the lower performance of the consumer-grade device with regard to FRR, obtaining a FAR of 1% shows feasibility, as this threshold is applicable for low security applications (see Section 5.2.3). The FRR of 22.4% is close to the upper threshold of 20% described in [19] for real-world systems. Additionally, previous studies have shown that users value an authenticator being quick and effortless as more important than its being accurate in terms of false rejects [50]. Therefore, the result margins are promising, though it remains to be tested if this FRR is acceptable for users of a brainwave system and how it can be improved. Even if there is a sacrifice in FRR, the low-cost setting has higher potential of being leveraged in real-world applications due to the BCI headsets being lightweight, wireless, and easier to setup.

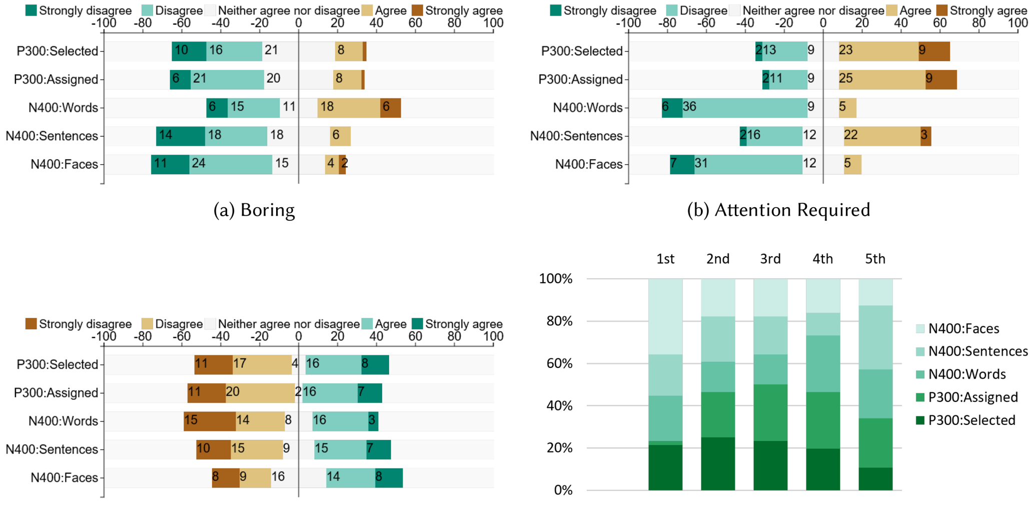
6 USABILITY

This section describes our user study to evaluate usability aspects, reporting quantitative and qualitative results.

6.1 User Study Design and Methods

Design. Each person taking part in the overall authentication experiment was asked to fill out a usability questionnaire (see Table 12 in Appendix A), which includes three categories of questions. First, we explore the perceived *usability of the five authentication tasks* asking if they are boring, require attention, and are appealing to repeatability on a daily basis (Q1-Q4). These questions are taken from Chuang's et al. work [22], though we ask for ratings on a 5-point Likert scale to allow for more granularity in the responses. Second, also on a 5-point Likert scale, we question about *device usability*, considering two dimensions: ability to set up the device (Q5) and overall usage experience (Q6). Third, we target *acceptance*. Inspired by the work of Payne et al. [65] on the acceptance of tokens as authenticators, we include two open-ended questions about potential problems (Q7) and suggestions for improvement (Q8) of the brainwave authentication concept. Note that users do not evaluate a prototype but the proposed authentication tasks and the perception of how a hypothetical brainwave-based system built on these tasks would work for them in daily life. The nature of the study is therefore exploratory and oriented to inform prototype design, whose evaluation would require further testing (see Section 7.4).

Analysis. Closed-ended usability questions elicited responses on 5-point Likert scales that we analysed using targeted hypothesis testing with $\alpha = .05$, selecting the appropriate test based on the data type and number of experimental conditions. We used the Friedman test for omnibus comparisons. Post hoc analysis with Wilcoxon signed-rank tests were conducted with a Bonferroni correction applied, to determine which authentication tasks differed significantly. As for the open-ended questions on user acceptance, we analyzed the responses following an iterative, inductive coding approach [54]. One member of the research team read responses and created the codebook



(a) Boring

(b) Attention Required

(c) Repeatable on a daily basis

(d) Enjoyability Ranking

Fig. 9. Participant answers to the statements: (a) *"The task was boring"*, (b) *"The task required a lot of attention"*, and (c) *"I could imagine to perform this task on a daily basis at a PC for authenticating"*, for the five implemented authentication tasks. Sub-figure (d) shows how respondents ranked the tasks depending on enjoyability.

with thematic codes (see Appendix A), and a second researcher independently coded the full set of data. We calculated the Cohen's Kappa (κ), a commonly used statistic reflecting agreement among coders, which corrects for how often ratings might agree by chance [23]. The inter-coder reliability for the final codes was satisfactory for both questions[13]: excellent agreement for Q7 on envisioned problems ($\kappa = 0.91$) and substantial for Q8 on suggested improvements ($\kappa = 0.76$). The cases where the coders differed in their final codes were discussed and reconciled.

6.2 Results

All 56 subjects replied to the Likert-ranked questions about the usability of authentication tasks and device. With regard to the open-ended questions, 28 subjects named potential problems, and 45 reported improvement suggestions for a brainwave authentication system. Here we analyze these data, providing representative user quotes when meaningful.

6.2.1 Perceived Usability. **Usability of the Authentication Tasks.** The graphs in Figure 9 show participants' answers about tasks' usability. Answers to "boring" and "required attention" were coded from Strongly Agree (SA) = 1 to Strongly Disagree (SD) = 5, and answers to "Repeatability", from SD = 1 to SA = 5. Therefore, higher values always indicate more positive evaluations.

Analyzing the responses regarding *boredom*, protocols were rated differently ($\chi^2(4) = 108.864$, p < .05). More specifically, there were statistically significant differences (p < .01) in all cases except between the P300:Assigned and P300:Selected, and the N400:Sentences and N400:Faces. The N400:Words protocol received the lowest grades with a median rating of 3 ($\mu = 2.95$, $\sigma = 1.21$). With slightly better grades, the P300:Selected ($\mu = 3.46$, $\sigma = 1$) and P300:Assigned ($\mu = 3.39$, $\sigma = 0.93$), received a median of 3 and present no statistically significant differences. At the other extreme, the N400:Faces protocol ($\mu = 3.78$, $\sigma = 0.99$), and the N400:Sentences ($\mu = 3.71$, $\sigma = 0.97$), with the same median rating of 4 and no statistically significant difference, got the best evaluations. About

[13] Generally, κ values of 0.4 to 0.75 are considered moderate to good, and values of >0.75 represent excellent agreement [77].

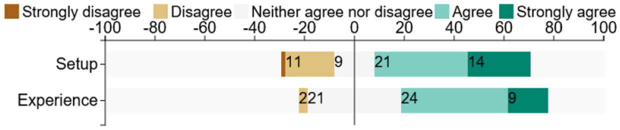


Fig. 10. Participant responses to the statements *"I could imagine to put the headset on myself after a short introduction"* (Setup) and *"My experience with the headset was very positive"* (Experience).

the latter, one of its positive aspects is that the sentences were unexpected and sometimes funny, which makes the task more engaging, as this participant put it in the open-ended answers:

> *"I like the idea with incongruent sentences. Generally, I think that it is important to include something funny or encouraging to avoid boredom"*. (P28)

When it comes to *required attention*, tasks were also rated differently ($\chi^2(4) = 158.501$, $p < .05$). Statistically significant differences ($p < .01$) appear in all cases except between the P300 protocols and the pair N400:Faces-N400:Words. The protocols with lower grades are the P300:Assigned (μ = 2.5, σ = 1.09) and the P300:Selected (μ = 2.57, σ = 1.13), both with a median of 2 and no statistically significant differences. Participants rated the attention demand of the N400:Sentences task (μ = 2.85, σ = 1.03) slightly better, with a median of 3. But the highest rates were assigned to N400:Faces (μ = 3.73, σ = 0.8) and N400:Words (μ = 3.77, σ = 0.76), both with a median of 4 and no statistically significant differences.

The responses regarding *envisioned daily usage* show differences too ($\chi^2(4) = 62.254$, $p < .05$), but they exhibit a smaller variance compared to the prior questions. In this case, N400:Faces (μ = 3.09, σ = 1.27), with a median of 3, is the best rated task. In turn, the N400:Words (μ = 2.61, σ = 1.3) got the worst evaluation, with a median of 2. The rest of the authentication tasks fall in the middle. Statistically significant differences ($p < .01$) appear in all cases except between the P300 protocols, and between P300 and N400:Sentences.

Finally, when we asked participants to rank the authentication tasks, the most enjoyable protocol was the N400 Faces, chosen by 36% (20) of the respondents. At the other end of the rank, the N400:Sentences task was selected as the least enjoyable by 30% (17) of the participants. Overall, image-based tasks are preferred over text-based ones, as it was also recalled by several participants in the open-ended questions:

> *"Picture recognition is better than text recognition"*. (P22)

Usability of the EEG Device. As it can be seen in Figure 10, most of the participants (62.5%) think they will be able to put on the headset by themselves, while only a 21.5% (12) reported that they do not imagine themselves completing the device setup. A plausible reason for this 21.5% could be that the headset setup required several minutes in some cases, where the hair density between the electrodes and the skin was thick. Nevertheless, the experience using the headset was mostly rated positive, with a 59% (33) of participants agreeing or strongly agreeing to this perception and no reported strong disagreements.

These results indicate that authentication using the EPOC+ headset could be accepted (positive experience) but the usability of the device can still improve. In this sense, as we will see in more detail when discussing the open-ended questions, responses like: *"simplify the headset"*, *"not so many contact points, easier self-employed setup"*, showed the importance of device simplicity. In the end, the headset seemed to be acceptable for a prototype, but for day-to-day use, subjects were emphasizing the need for an easier device.

6.2.2 Attitudes Towards Acceptance. Answering the two open-ended questions, the participants had several ideas of potential problems and suggestions for improvement.

Problems. Participants identified issues related to the *brainwaves* (28%), the *device* (22%), and the overall authentication *system* (50%). First, users reported concerns about the uniqueness of brainwaves and their stability against e.g., emotional influences due to stress or illness. They were also worried that familiarization with the stimuli would result in weaker brainwave responses and lead to authentication errors. Besides, one subject wondered if not being fully attentive, or as he/she put it *"having meandering thoughts"*, would affect authentication. Second, the negative points about the device were the cost, its design, and the complex setup process. Similarly, users highlighted the technical problems, such as the imprecision of the sensors. Third, participants criticized aspects of the system as a whole, specially its performance (authentication speed), usability, and the level of security and privacy provided. As illustrated by the following sample answers, users are worried about the strength of this type of authentication against attacks (even mind manipulation) and about the usage of brainwaves to infer sensitive personal information.

> *"Skepticism of the user regarding data security and other aspects which could be figured out about the users, which the user does not want."*. (P9)
>
> *"Changing of individual opinion due to presented stimuli, e.g., in particular politicians"*. (P41)

In the usability category, the inclusiveness of the brainwave authentication system was the most frequent topic. Participants remarked that using sentences as stimuli would not work to authenticate children and that the system might not be usable for people with different cognitive abilities.

Suggestions for improving. Participants reported ideas that fall in three categories: *device* improvements (18%), *protocol* improvements (39%), and *system* improvements (42%). Regarding the device, users pointed to different designs that blend more naturally with everyday life, such as integrating EEG readers within headphones or hats. Another frequent comment was the need to reduce the number of electrodes and make the device simpler and easy to handle. Regarding the improvement of protocols, subjects expressed a preference for visual stimuli vs textual stimuli and call for authentication tasks that are enjoyable or "cool". As alternative tasks, for example, two participants mentioned that they *"would be interested in authentication using music or tones"*. In the last category of suggestions, targeting the overall system, performance was the most frequent concern. Users suggest to *"Keep the authentication process as short as possible"*, because otherwise *"one sees the repeated, three second long typing of a password as less annoying than performing one of these [brainwave authentication] tasks as a whole"*. The effort, as stated by one of the respondents *"needs to be adapted to the required security level"*.

7 DISCUSSION

Here we report lessons learned when designing protocols for brainwave authentication, report security considerations, and discuss practical implementation aspects and limitations.

7.1 Protocol Design

Design Effort. We argued in Section 5 that one potential reason influencing the performance and comparability of the authentication protocols was the different available number of samples for training the models, which, in our study, was affected by the protocol design effort. The number of epochs usable for classification is limited by the total number of target stimuli, i.e., those that generate an ERP, presented during the experiment. As summarized in Table 6, both the N400:Sentences and N400:Faces have less total stimuli in comparison to their counterparts. There are two reasons for this: highest elicitation effort (more time required for stimuli presentation) and low stimuli

Table 6. Design Aspects of Brainwave Acquisition Protocols

Design Aspects	**P300:Selected**	**P300:Assigned**	**N400:Words**	**N400:Sentences**	**N400:Faces**
Avg. time[a] between target stimuli (s)	6	6	4.15[b]	14	6
# Target stimuli per round	6	6	13	6	10
# Protocol rounds	3	3	3	1	1

[a] Rounded.

[b] Plus the duration of the preceding priming video (24s in our experiment).

Table 7. Overall Comparison of Authentication Protocols

Criteria	**P300:Selected**	**P300:Assigned**	**N400:Words**	**N400:Sentences**	**N400:Faces**
Boredom	+	+	-	+ +	+ +
Required level of attention	- -	- -	+ +	-	+ +
Daily Usage	-	-	- -	-	+
Enjoyability	+	-	+	- -	+ +
Elicitation effort	+ +	+ +	+	- -	- -
Stimulus reusability	+ +	+ +	+	- -	- -

reusability. While it is rather quick to present new stimuli in the N400:Words, N400:Faces, and P300 protocols, that was not possible in the N400:Sentences. In this case, the subjects first had to be primed on the congruent form of a sentence and then later on shown the incongruent version to obtain the desired ERP in response. This process takes about 14 seconds per sentence in total, which results in a smaller number of stimuli per minute. Furthermore, the incongruent sentences need to be altered each time, otherwise they would not appear incongruent to the users anymore after a small number of iterations. Similarly, the N400:Faces also suffers from this effect, i.e., an unknown face would not lead to the same reaction if it was shown repeatedly. Because of the lack of stimuli reusability, we limited the execution of these protocols to just one round in our experimental setting, with the consequential decrease in the number of samples. In the N400:Words protocol, a video and the associated words can be used several times, since only the interaction between the words and the video are important. Stimulus familiarity poses a challenge to practical implementations, which need to scalably construct a corpus of stimuli with enough variation. While practicality remains to be thoroughly analysed, recent advances in generative AI models for text to image/video and vice-versa could aid in this task [69]. Stimuli creation could be assisted by machine learning models to automatically generate synthetic faces, or to generate videos and words related to them.

Overall, the best design case is that of the P300 protocols. Here, the stimuli can be endlessly reused because the brain reaction responds to an infrequent event, the oddball, but it is not related to the semantic processing and so is unaltered by stimulus familiarity.

Overall Protocol Comparison. We provide a comparative summary of the analyzed protocols to inform the design of future brainwave authentication systems (see Table 7).

Considering classification performance, the P300 tasks and the N400:Words are the best options, but closely followed by the other options, whose performance will potentially increase with a higher number of samples. Regarding usability, the appeal of P300 tasks could be improved to gain acceptability in real-world implementations. On the one hand, we observe that usability improves when users select their own secret image. This preference on active selection was also observed by Chuang et al. [22] in protocols where users either had to choose or were imposed a mental task for authentication. On the other hand, looking at the best performing task with regard to usability, the N400:Faces, it could be interesting to explore if using faces stimuli in P300 tasks can help improve users attitudes. The most positive of P300 protocols is that they are the easiest to implement, which might be a reason why research with ERPs so far mostly focused on these potentials.

The main negative aspect in N400 tasks is the complexity of the protocol design. Thus, research towards facilitating this design process is desirable.

7.2 Security

In this paper we covered a zero effort attacker model, but, like in other biometric methods, adversaries can also attack brainwave authentication by compromising different parts of the system [9]. The most applicable attack vector that targets specific users is arguably the *replay attack*, where the adversary injects a previously recorded sample of the biometric. Furthermore, with the current advance of machine learning techniques, it is also possible to generate fake brainwave data using Generative Adversarial Networks [67]. In this regard, if the authentication stimuli vary for each authentication attempt (order, type), the elicited brain responses will vary accordingly, but still provide the required user-specific features. This type of challenge-response protocol implies that the attacker should be able to output results interactively in real-time, as the stimuli are not known in advance, which makes the attack harder to implement. Furthermore, an attacker observing a user while authenticating learns nothing about the brainwaves. Mimicry attacks, which are feasible for other biometrics (voice, gait), are not applicable because the adversary cannot imitate non-volitional user responses.

The acquisition of EEG signals also raises privacy issues because brainwaves correlate e.g., with our mental states, cognitive abilities, and medical conditions [75]. An adversary that controls the authentication stimuli, such as an honest-but-curious authentication provider, could manipulate them to infer private data. Martinovic et al. [51] demonstrated the feasibility of this type of attack. They successfully proved that, by manipulating visual stimuli, EEG signals could reveal users' private information about their bank cards, PIN numbers, area of living, and if they knew a particular person. Frank et al. [31] go even further, showing that it is possible to extract private data from EEG recordings using subliminal stimuli (short duration images embedded in visual content) that cannot even be consciously detected by users.

With the potential wide adoption of BCI applications in our everyday lives, security and privacy concerns are rising [10, 14]. Our user study and other previous research [53] show that users are concerned about 'mind reading', but some people are already giving their brainwaves to third parties that offer brain-controlled games or relaxation applications. It is therefore paramount to research the security and privacy implications of using brainwaves in computer systems and work to design appropriate countermeasures before mainstream adoption.

7.3 Practical Implementation Aspects

Time to authenticate. A prototype implementation based on the P300:Selected brainwave authentication algorithm would require an initial *enrollment phase*. This means approximately 1 minute of brain data recording while the user looks at images in their PC. This phase could be extended to collect a higher amount of samples for training the system and broken into several shorter sessions for user convenience. It would be useful to implement a sample quality detector to adapt the duration of the enrollment process, similar to how fingerprint systems ask the user to place the finger in different angles until enough data is gathered for successful operation. Next, the *authentication phase* would require a minimum of 6 seconds to authenticate the user. Upon unsuccessful authentication attempts, the trial could be repeated, but this would also result in a higher FAR. Fallback mechanisms should be implemented in case the authentication does not succeed in a reasonable time and with sufficient security guarantees. Based on previous empirical research [82], the average time to authenticate with 8-character random passwords is around 7.5 seconds (12.8–13.2 seconds in tablet/smartphones [82]). Therefore, brainwave authentication is better in a best-case execution. But even if it takes longer, it has to be considered that usability perceptions can deviate from objective performance measures. For example, research shows

Table 8. Comparison of P300: Selected Brainwave Authentication Against Passwords and Fingerprint using Bonneau et al.'s framework [15]

	Usability								Deployability						Security										
Scheme	Memorywise-Effortless	Scalable-for-Users	Nothing-to-Carry	Physically Effortless	Easy-to-Learn	Efficient-to-Use	Infrequent-Errors	Easy-Recovery-from-Loss	Accessible	Negligible-Cost-per-User	Server-Compatible	Browser-Compatible	Mature	Non-Proprietary	Resilient-to-Observation	Resilient-to-Targeted-Impersonation	Resilient-to-Throttled-Guessing	Resilient-to-Unthrottled-Guessing	Resilient-to-Internal-Observation	Resilient-to-Leaks-from-Other-Verifiers	Resilient-to-Phishing	Resilient-to-Theft	No-Trusted-Third-Party	Requiring-Explicit-Consent	Unlinkable
Passwords			●		●	●	○	●	●	●	●	●	●	●		○						●	●	●	●
Fingerprint	●	●	●	○	●	●	●		○				●		○		●						●	●	
Brainwaves	●	●		●	●	○		○	○						●	●	●		●	●	●	●	●	●	

"●" indicates that the scheme provides the benefit; and "○" means that the benefit is somewhat provided.

evidence that graphical authentication schemes are perceived as more joyful than passwords even if the login time may exceed that of passwords [47, 83]. In this sense, the N400:Faces is promising given the positive ratings on enjoyability obtained in the user study.

Extended Comparison. We use the framework of Bonneau et al. [15] to compare brainwave authentication against passwords (the most common solution) and fingerprint (the most used biometric). Table 8 summarizes this comparison according to the 25 criteria provided by the framework, grouped in *usability*, *deployability*, and *security* benefits. It can be seen that brainwave authentication provides better usability than passwords, and it could be comparable in the future to that of fingerprints if FRRs improve significantly. Still, they have the disadvantage of having to carry a device. A potential improvement with regard to fingerprint usability is that brainwave biometrics have the potential to be revoked. As demonstrated by Lin et al. [44], the EEG response changes with presented stimuli, and they are sufficiently different to provide revocability. However, future work should thoroughly test revocability for the concrete tasks we evaluated. Accordingly, we only consider the support of this benefit as partially achieved, backed by existing work. Overall, usability is not a one-size-fits-all, and the applicability of brainwaves should be evaluated for different use-cases.

On the security criteria, brainwaves bring additional benefits because they are not observable and cannot be mimicked. Targeted impersonation attacks with synthetic or replayed data can be countered using the challenge(stimulus)-response nature of the brainwave authentication protocol. This allows the system to check response freshness and whether reactions correspond to stimuli that are meaningful for the legitimate user. Furthermore, as the adversary would need to interact with a legitimate authentication provider to obtain those per-user stimuli, we get resilience to phishing. The main security challenge is to reduce the FAR. Besides, brainwaves have the worst deployability, though the evaluation framework criteria focus on applicability to web authentication. Aspects like browser compatibility could be addressed by implementing brainwave authentication as part of the FIDO/WebAuthn protocols [84], currently supported in modern browsers. Additionally, there are other domains and use-cases outside the web realm where brainwaves could become practical.

Use-cases. The proposed brainwave authentication system was conceptualized for accessing PC applications, but the stimuli can be easily adapted to other devices and scenarios. Furthermore, once authenticated with the brainwave protocol, the user continues to have measurable brain activity, which can be leveraged for continuous authentication while wearing the headset. Brainwaves can be practical when users already wear an EEG reader for another application and a keyboard is inconvenient/unavailable. For example, authentication in **Virtual Reality (VR)** applications is still challenging as passwords are clearly unpractical. But modern VR headsets are introducing EEG sensors, making them a perfect scenario to apply our mechanisms. Additionally, with the ongoing miniaturization and integration of EEG sensors in devices that people commonly use (e.g., earbuds), having to carry them can be less problematic. Moreover, brainwaves could be augmented with other sensors that collect implicit biometrics (e.g., eye gaze) to improve authentication accuracy and, therefore, increase security.

7.4 Limitations

We acquired brainwaves in a lab environment and during a single recording session so we could not evaluate reliability and robustness against potential variations of brain reactions across multiple sessions or with regard to noise or changing conditions. Nevertheless, we expect our system to be robust given the good level of permanence of EEG distinctive features, as demonstrated in previous research. Maiorana et al. [48] explored the stability of brainwaves for authentication across three sessions, with a maximum separation of around one month, concluding that there was no evident variability trend that impacts the recognition results. Furthermore, ERPs are less sensitive to background noise than continuous EEGs and, even if latency/amplitude might vary with external factors like stress, tiredness, and the like [20], ERPs reflect morphological components (e.g., skull thickness) that are more stable [5, 13]. Additional experiments in real-life conditions should be conducted to validate this hypothesis. In our experiments, we observed a high variability in the performance of different brainwave authentication tasks. We speculate that the number and quality of registered samples impacts the results, but further research is required to understand the factors inducing this variability and how to reduce their effect. It would be valuable to investigate the scalability of the results to larger populations.

Our user study on usability is based on a limited sample of the population, mostly young and technically-savvy users. Bigger and diverse sets of users would yield a more comprehensive picture of the usability issues in brainwave authentication. We described the system to our participants embedding it in a realistic use case: we told them that they would have to watch one task out of the set of tasks in the experiment once a day, and this would replace the need to type passwords for their applications. With this description, a perfect implementation is assumed. Another limitation is that we rely on self-reported qualitative feedback about intended future behavior based on participants perception of the described system, which might not accurately reflect reality [43]. With this study design, our goal is to describe problems that could hinder the adoption of brainwave-based authentication to consider when designing actual prototypes or experiments, but we do not claim any generalizable findings. To achieve ecological validity, we need to evaluate the actual usability of authentication prototypes in real scenarios, applying established metrics in authentication research [25, 70] that were not suitable to be applied in this early stage of research, including the **Standard Usability Scale (SUS)** [17], speed measurements, and error rates.

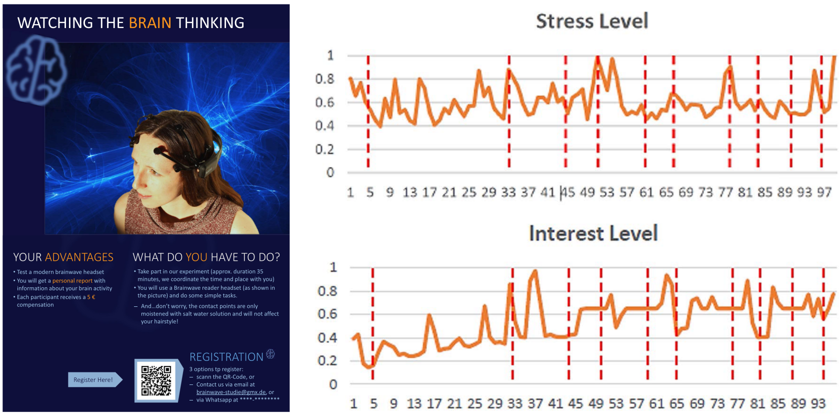
8 CONCLUSION

We contribute to the literature on biometrics with the first comparative study on the usability and performance of brainwave authentication protocols based on endogenous Event Related Potentials using consumer-grade EEG readers. Our results show the feasibility of authentication by recording brain activity while users are exposed to short sequences of stimuli (images/words). With regard to perceived usability, users are positive about this type of system but call for simpler headsets and fast authentication times. Considering participants feedback, we highlight the need to conduct extensive privacy research before brainwave-based applications become mainstream. When contextualizing our results, we found out that comparability with other works is hampered by differences in experimental conditions and performance reporting schemes, but also because the sample sizes used in the literature are very small (the majority ≤ 10). We therefore contribute our dataset to improve the availability of samples and provide a source for common benchmarking. To bridge the comparability gap, the authentication community should strive to establish a consistent approach for communicating performance metrics.

A APPENDIX: ADDITIONAL MATERIAL

Open Data. The anonymized dataset and experiment material are available at https://git.scc.kit.edu/kr2925/brainwave-authentication.

Recruitment Material. Figure 11(a) shows the study advertisement, distributed as flyers and posters across the university campus and announced during lectures.



WATCHING THE BRAIN THINKING

YOUR ADVANTAGES

- Test a modern brainwave headset
- You will get a personal report with information about your brain activity
- Each participant receives a 5 € compensation

WHAT DO YOU HAVE TO DO?

- Take part in our experiment (approx. duration 35 minutes, we coordinate the time and place with you)
- You will use a Brainwave reader headset (as shown in the picture) and do some simple tasks.
  – And...don't worry, the contact points are only moistened with salt water solution and will not affect your hairstyle!

Register Here!

REGISTRATION

3 options tp register:
– scann the QR-Code, or
– Contact us via email at brainwave-studie@gmx.de, or
– via Whatsapp at ****-********

(a) Recruitment flyer (b) Personal Report

Fig. 11. (a) Flyer for recruitment. (b) Snapshot of the graphs for "Stress Level" and "Interest Level" (measured every 10 seconds) provided in the personal brainwaves report given to participants after the study. The vertical bars signal task changes.

Experiment Instruments. The experiment starts with the experimenter providing the participant with a consent form. After giving consent, the participant is shown a sequence of images (printed in paper) and told to choose one. He/She is also assigned another picture and told to remember it. Next, the participant sits in front of a PC screen, receives a paper form to write answers in the subsequent steps, and the experimenter fits the EEG headset to him/her. From that moment on, the experimenter tells the participant to follow the instructions in the screen, summarized in Tables 9, 10, and 11. Once the brainwave collection is finished, the participant is asked to fill a paper survey to evaluate the perceived usability of a brainwave authentication system based on the performed tasks and gather demographic data. The survey questionnaire is detailed in Tables 12–14 contain the codebook used to analyse free text questions. After the survey, once the experiment is finished, the participants get their compensation and have the chance to ask questions. They will be contacted in the following days to receive a personal report on their brain activity during the experiment.

Table 9. Brainwave Collection Experiment Instructions

Introduction

Welcome to our Brainwave study. The study will take approximately 30 minutes. Please take a seat now and try to move as little as possible during the tasks. You can move during breaks. When you're ready to start, please press the space bar. You can use the space bar to navigate to the next step in the entire process. Let's begin. **Baseline**

Now keep your eyes open for 20 seconds and relax. If possible, try not to blink during the 20 seconds.

Now relax with your eyes closed for 20 seconds. Keep your eyes closed until you hear an acoustic signal.

Now open your eyes and press the space bar to start.

P300:Selected and P300:Assigned

You will see the following task a total of six times during the experiment. You will see a series of pictures during the task. Press the space bar to start the task.

Now remember the picture you selected (were assigned) at the beginning of the experiment. Your task in the following is to count how often exactly this picture occurs. Press the space bar to start.

[Images]

How often have you recognized your picture? Write the number in the space provided on your paper.

N400:Words

You will now watch a video. After the video, you will be asked to note three terms you associate with the video. Press the space bar to start the video now.

The video is about to start. Watch carefully.

[Video]

Please write down three terms you associate with the video in the space provided on your paper. Press the space bar to continue.

You will now see a series of words. Read carefully. Press the space bar to start the series of words.

Subliminal Video

You will now watch another video. Watch carefully. Press the space bar to continue.

[Video]

N400:Sentences

Next you will be shown individual words. These result in sentences. Read carefully and try to visualize the sentences. Press the space bar to continue.

[Sentences]

N400:Faces

You will now see some more pictures. Watch carefully. Press the space bar to start.

[Face Images]

End of Experiment

Thank you very much for participating in our experiment! Please contact your experimenter now. She will conduct a small final survey with you.

After-tasks

Thank you, you have completed [Task i] out of [N] tasks.

Personal Report. The report explains the different type of brainwaves a person has in different states (e.g., when attentive or idling), describing where they originate and which electrodes capture them. It also provides graphs showing the mental state of the participant during the experiment as derived from his/her brain activity. The graphs show: stress level, interest level, engagement level, relaxation level, focus level, and excitement level. Figure 11(b) shows a partial example of the graphs included in the personal report.

Table 10. Words used in the N400:Words Authentication Task, Related and Unrelated to a Video Showing Driving Cars

Related	Car, Track, Road, Highway, Vehicle, Speed, Steering Wheel, Toll, Expressway, Sports car, Automobile, Driver
Unrelated	Apple, Biology, Moon, Circle, Kitchen, Hunger, Opera, Mushroom, Hare, Price, Hotel, Ladder, Selection, Hairstyle, Studies, Chalk, Producer

Table 11. Sentences used in the N400: Sentences Task

Sentences	**Priming (Probing) Ends**
I drink coffee with milk and	sugar(socks)
Ted smiled and bit his bottom	lip(rainbow)
The prison ward walked along the	row(moon)
A horse has thrown a	shoe(plane)
Steve sat down to eat his	lunch(car)
He put the fork on the	table(door)

Table 12. Usability and Demographic Questions

Introduction

We want to build an authentication system based on brainwaves. In order to use such a system, you would have to watch one task out of the set of tasks in the experiment once a day. This step would replace all passwords for all applications you are currently entering. Please score the tasks with regard to their usage in a brainwave authentication system.

Perceived Usability of the Authentication Tasks

Please score the tasks based on three criteria. (1 = Strongly Agree to 5 = Strongly Disagree)
Q1. The task was boring
Q2. The task required a lot of attention
Q3. I could imagine to perform this task on a daily basis at a PC for authenticating
Q4. Please sort the tasks depending on how enjoyable they were (1 = Most Enjoyable, 5 = Least Enjoyable)

Perceived Usability of the Device

Q5. I could imagine to put the headset on myself after a short introduction (5 = Strongly Agree, 1 = Strongly Disagree)
Q6. My experience with the headset was very positive (5 = Strongly Agree, 1 = Strongly Disagree)

Acceptance

Q7. Do you envision any problems with an authentication system using these techniques?
Q8. Do you have any suggestion for designing an authentication system based on these techniques?

Demographics and Personal Information

Q9. Please indicate your gender. (Options: Male, Female, Other)
Q10. Please indicate your age. (Options: 18–24, 25–31, 32–38, 39–45, 46–52, 53–59, 60 and older)
Q11. Which hand is your dominant hand? (Options: Left, Right)
Q12. I felt rather stressed out during the last week. (5 = Strongly Agree, 1 = Strongly Disagree)
Q13. I feel tired today. (5 = Strongly Agree, 1 = Strongly Disagree)
Q14. Did you drink alcohol yesterday? (Options: Yes, No)
Q15. Did you consume caffeine during the last 12 hours? (Options: Yes, No)

Table 13. Categories and Codes used to Code Free Text Answers on Envisioned Problems of Brainwave-based Authentication

Category	Codes	Definitions	Examples
Device (n=16)	Design (1%)	Participants report problems in the EEG headset design	“The headsets would need to be smaller, so that it would be practical to take it anywhere”
	Setup (5%)	Participants report problems with the EEG headset setup	“It is too complicated to put on the headset self-employed”
	Cost (2%)	Participants identify the price of the EEG headset as a problem for adoption	“Procurement to expensive”
	Technical Problems (12%)	Participants report envisioned technical problems with the EEG headset operation	“The electrodes are not functioning properly, the system is very sensitive to movements”
Brainwaves (n=21)	Stability (20%)	Participants report concerns with the stability of brainwaves with external (e.g., noise) and internal conditions (e.g., mental states like being tired or under stress)	“Under stress brainwaves are maybe different?”
	Uniqueness (8%)	Participants report concerns with the uniqueness of their brainwaves	“I think it is much harder to get evidence for the uniqueness of individual brainwaves for unambiguous identification than with fingerprint genes”
System (n=37)	Performance (9%)	Participants report system performance, in terms of time to authenticate, as a problem in brainwave based authentication systems	“It takes too long to perform this every day”
	Usability (16%)	Participants report usability problems -other than time performance- of the overall authentication system.	“An authentication system using sentences could be problematic for some people, like for example kids”
	Security & Privacy (13%)	Participants report problems related to the security and privacy aspects of a brainwave authentication system	“With improved and advanced technology: contact-less brainwave reading maybe possible at some point and then technical imitation”
	Deployment (6%)	Participants identify problems that arise when the system is deployed in real-life scenarios	“The results of the recording would need to be imported into a new single-sign-on system which also is not straightforward.”
	Technical Problems (4%)	Participants report envisioned technical problems in the overall brainwave-based authentication system	“Problems during analysis”

Percentages in parentheses indicate the number of times a code was used.

Table 14. Categories and Codes used to Code Free Text Answers on Suggestions to Improve Brainwave-based Authentication

Category	Codes	Definitions	Examples
Device (n=7)	Design (13%)	Participants suggest concrete changes in the EEG reader design, e.g., to modify its shape	“As a hat”
	Simplicity (5%)	Participants point at the general need to simplify the EEG recording process, without giving concrete suggestions on how to do it	“Simplify the headset”
Protocol (n=15)	Design (31%)	Participants suggest modifications to the authentication tasks, or point at features in the tested tasks considered desirable that should be included in a brainwave authentication system	“I would be interested in authentication using music or tones”
	Enjoyability (7%)	Participants report identify as positive that the authentication tasks are pleasant	“I like the idea with incongruent sentences. Generally, I think that it is important to include something funny or encouraging to avoid boredom”
System (n=16)	Performance (21%)	Participants report that a brainwave based authentication system should have a good performance in terms of time to authenticate	“The duration of the authentication has to be kept as short as possible.”
	Deployment (7%)	Participants report potential applications of brainwave-based authentication, or identify required improvements/adaptations of the system when deployed in real-life scenarios	“For securing the entry to buildings”
	Usability (13%)	Participants suggest to improve usability aspects -other than time performance- of the overall authentication system	“Less effort for an integration into everyday life" ”

Percentages in parentheses indicate the number of times a code was used.

REFERENCES

[1] Sherif Nagib Abbas, Mohammed Abo-Zahhad, and Sabah Mohammed Ahmed. 2015. State-of-the-art methods and future perspectives for personal recognition based on electroencephalogram signals. *IET Biometrics* 4, 3 (September 2015), 179–190.

[2] M. Abo-Zahhad, Sabah M. Ahmed, and Sherif N. Abbas. 2016. A new multi-level approach to EEG based human authentication using eye blinking. *Pattern Recognition Letters* 82 (2016), 216–225.

[3] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (December 1974), 716–723.

[4] Patricia Arias-Cabarcos, Thilo Habrich, Karen Becker, Christian Becker, and Thorsten Strufe. 2021. Inexpensive Brainwave Authentication: New Techniques and Insights on User Acceptance.

[5] Blair C. Armstrong, Maria V. Ruiz-Blondet, Negin Khalifian, Kenneth J. Kurtz, Zhanpeng Jin, and Sarah Laszlo. 2015. Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics. *Neurocomputing* 166 (2015), 59–67.

[6] Corey Ashby, Amit Bhatia, Francesco Tenore, and Jacob Vogelstein. 2011. Low-cost electroencephalogram (EEG) based authentication. 442–445.

[7] Michael P. Barham, Gillian M. Clark, Melissa J. Hayden, Peter G. Enticott, Russell Conduit, and Jarrad A. G. Lum. 2017. Acquiring research-grade ERPs on a shoestring budget: A comparison of a modified emotiv and commercial SynAmps EEG system. *Psychophysiology* 54, 9 (2017), 1393–1404.

[8] Md. Khayrul Bashar, Ishio Chiaki, and Hiroaki Yoshida. 2016. Human identification from brain EEG signals using advanced machine learning method EEG-based biometrics. 475–479.

[9] Karen Becker, Patricia Arias-Cabarcos, Thilo Habrich, and Christian Becker. 2019. Poster: Towards a Framework for Assessing Vulnerabilities of Brainwave Authentication Systems. 2577–2579.

[10] Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, and Sasitharan Balasubramaniam. 2019. Cybersecurity in brain-computer interfaces: State-of-the-art, opportunities, and future challenges. *arXiv:1908.03536* (2019).

[11] Niels Birbaumer and Leonardo G. Cohen. 2007. Brain–computer interfaces: Communication and restoration of movement in paralysis. *The Journal of Physiology* 579, 3 (2007), 621–636.

[12] D. H. R. Blackwood and W. J. Muir. 1990. Cognitive brain potentials and their application. *British Journal of Psychiatry* 157, S9 (December 1990), 96–101.

[13] Maria V. Ruiz Blondet, Sarah Laszlo, and Zhanpeng Jin. 2015. Assessment of permanence of non-volitional EEG brainwaves as a biometric. 6 pages.

[14] Tamara Bonaci, Ryan Calo, and Howard Jay Chizeck. 2014. App stores for the brain: Privacy & security in Brain-Computer Interfaces. 7 pages.

[15] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes., 553–567 pages.

[16] Katharine Brigham and B. V. K. Vijaya Kumar. 2010. Subject identification from electroencephalogram (EEG) signals during imagined speech. 8 pages.

[17] John Brooke, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L McClelland. 1996. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[18] Daniel Buschek, Alexander De Luca, and Florian Alt. 2015. Improving accuracy, applicability and usability of keystroke biometrics on mobile touchscreen devices. 1393–1402.

[19] Ann Cavoukian and Alex Stoianov. 2007. *Biometric Encryption: A Positive-sum Technology that Achieves Strong Authentication, Security and Privacy*. Information and Privacy Commissioner, Ontario.

[20] Hui-Ling Chan, Po-Chih Kuo, Chia-Yi Cheng, and Yong-Sheng Chen. 2018. Challenges and future perspectives on electroencephalogram-based biometrics in person recognition. *Frontiers in Neuroinformatics* 12 (2018), 66.

[21] Gabriel Chuang and John Chuang. 2016. Passthoughts on the Go : Effect of Exercise on EEG Authentication. (2016). unpublished.

[22] John Chuang, Hamilton Nguyen, Charles Wang, and Benjamin Johnson. 2013. I think, therefore I am: Usability and security of authentication using brainwaves. *Lecture Notes in Computer Science* 7862 LNCS (2013), 1–16.

[23] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[24] Hieu Dao, Dinh-Huan Nguyen, and Minh-Triet Tran. 2021. Face Recognition in the Wild for Secure Authentication with Open Set Approach. 338–355.

[25] Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. 2014. Now you see me, now you don't: Protecting smartphone authentication from shoulder surfers. 2937–2946.

[26] Jacques B. Debruille, Jaime Pineda, and Bernard Renault. 1996. N400-like potentials elicited by faces and knowledge inhibition. *Cognitive Brain Research* 4, 2 (1996), 133–144.

[27] Matthieu Duvinage, Thierry Castermans, Mathieu Petieau, Thomas Hoellinger, Guy Cheron, and Thierry Dutoit. 2013. Performance of the emotiv epoc headset for P300-based applications. *Biomedical Engineering Online* 12, 1 (2013), 56.

[28] Emotiv Systems. Accessed: 31.07.2019. Emotiv EEG Headset Comparison Page, https://www.emotiv.com/comparison/.

[29] EU. Accessed: 30.04.2019. General Data Protection Regulation, https://gdpr-info.eu/.

[30] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[31] Mario Frank, Tiffany Hwu, Sakshi Jain, Robert Knight, Ivan Martinovic, Prateek Mittal, Daniele Perito, and Dawn Song. 2013. Subliminal probing for private information via EEG-based BCI devices. *CoRR* abs/1312.6052 (December 2013), 1–12.

[32] Qiong Gui, Maria V. Ruiz-Blondet, Sarah Laszlo, and Zhanpeng Jin. 2019. A survey on brain biometrics. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–38.

[33] Lindsay F. Haas. 2003. Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry* 74, 1 (2003), 9–9.

[34] InteraXon Inc. [n. d.]. https://choosemuse.com/. Accessed: 05.02.2020.

[35] Isuru Jayarathne, Michael Cohen, and Senaka Amarakeerthi. 2017. Survey of EEG-based biometric authentication. 324–329.

[36] Benjamin Johnson, Thomas Maillart, and John Chuang. 2014. My thoughts are not your thoughts. 1329–1338.

[37] Emily S. Kappenman, Jaclyn L. Farrens, Wendy Zhang, Andrew X. Stewart, and Steven J. Luck. 2021. ERP CORE: An open resource for human event-related potential research. *NeuroImage* 225 (2021), 117465.

[38] Juris Klonovs, Christoffer Kjeldgaard Petersen, Henning Olesen, and Allan Hammershoj. 2013. ID Proof on the Go: Development of a mobile EEG-based biometric authentication system. *IEEE Vehicular Technology Magazine* 8, 1 (2013), 81–89.

[39] Belal Korany, Chitra R. Karanam, Hong Cai, and Yasamin Mostofi. 2019. XModal-ID: Using WiFi for Through-Wall Person Identification from Candidate Video Footage. 15 pages.

[40] Marta Kutas and Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology* 62, 1 (2011), 621–647.

[41] Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207, 4427 (1980), 203–205.

[42] Marta Kutas and Steven A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307 (1984), 161–163.

[43] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction.* Morgan Kaufmann.

[44] Feng Lin, Kun Woo Cho, Chen Song, Wenyao Xu, and Zhanpeng Jin. 2018. Brain password: A secure and truly cancelable brain biometrics for smart headwear. 296–309.

[45] Fabien Lotte, L. Bougrain, A. Cichocki, M. Clerc, Marco Congedo, A. Rakotomamonjy, and F. Yger. 2018. A review of classification algorithms for EEG-based brain computer interfaces: A 10 year update. *Journal of Neural Engineering* 15, 3 (2018), 031005.

[46] Myndplay Ltd. [n. d.]. www.myndplay.com/. Accessed: 05.02.2020.

[47] Yao Ma and Jinjuan Feng. 2011. Evaluating usability of three authentication methods in web-based application. 81–88.

[48] Emanuele Maiorana, Daria La Rocca, and Patrizio Campisi. 2015. On the permanence of EEG signals for biometric recognition. *IEEE Transactions on Information Forensics and Security* 11, 1 (2015), 163–175.

[49] Anthony J. Mansfield and James L. Wayman. 2002. Best practices in testing and reporting performance of biometric devices. (2002).

[50] Shrirang Mare, Mary Baker, and Jeremy Gummeson. 2016. A study of authentication in daily life. In *Twelfth Symposium on usable Privacy and Security (SOUPS 2016)*. 189–206.

[51] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros, and Dawn Song. 2012. On the feasibility of side-channel attacks with brain-computer interfaces. 34–43.

[52] Takehiro Maruoka, Kenta Kambe, Hideki Harada, and Isao Nakanishi. 2017. A study on evoked potential by inaudible auditory stimulation toward continuous biometric authentication. 1171–1174.

[53] Nick Merrill, Max T. Curran, and John Chuang. 2017. Is the Future of Authenticity All In Our Heads? Moving Passthoughts From the Lab to the World. 10 pages. https://doi.org/10.1145/3171533.3171537

[54] Matthew B. Miles and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook.* Sage.

[55] Chisei Miyamoto, Sadanao Baba, and Isao Nakanishi. 2009. Biometric person authentication using new spectral features of electroencephalogram (EEG). 4 pages.

[56] Kusuma Mohanchandra. 2013. Using brain waves as new biometric feature for authenticating a computer user in real-time. *International Journal of Biometric and Bioinformatics* 7, 1 (2013), 49–57.

[57] Isao Nakanishi, Sadanao Baba, and Shigang Li. 2011. Evaluation of Brain Waves as Biometrics for Driver Authentication Using Simplified Driving Simulator. 71–76.

[58] Isao Nakanishi and Masashi Hattori. 2017. Biometric potential of brain waves evoked by invisible visual stimulation. 94–99.

[59] Isao Nakanishi and Takehiro Maruoka. 2019. Biometric authentication using evoked potentials stimulated by personal ultrasound. 365–368.

[60] Isao Nakanishi and Takuya Yoshikawa. 2015. Brain waves as unconscious biometrics towards continuous authentication - the effects of introducing PCA into feature extraction. 422–425.

[61] NeuroSky. Accessed: 30.04.2019. NeuroSky MindWave Family Description Page, http://neurosky.com/about-neurosky/.

[62] Ramaswamy Palaniappan. 2005. Multiple mental thought parametric classification: A new approach for individual identification. *International Journal of Signal Processing* 2, 4 (2005), 222–226.

[63] Ramaswamy Palaniappan. 2008. Two-stage biometric authentication method using thought activity brain waves. *International Journal of Neural Systems* 18, 01 (2008), 59–66.

[64] R. Palaniappan and D. P. Mandic. 2007. Biometrics from brain electrical activity: A machine learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 4 (2007), 738–742.

[65] Jeunese Payne, Graeme Jenkinson, Frank Stajano, M. Angela Sasse, and Max Spencer. 2016. Responsibility and tangible security: Towards a theory of user acceptance of security tokens. *arXiv preprint arXiv:1605.03478* (2016).

[66] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51, 1 (2019), 195–203.

[67] Tanya Piplani, Nick Merill, and John Chuang. 2018. Faking it, Making it: Fooling and Improving Brain-Based Authentication with Generative Adversarial Networks. 7 pages.

[68] Marios Poulos, Maria Rangoussi, and Nikolaos Alexandris. 1999. Neural network based person identification using EEG features. 1117–1120.

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[70] Scott Ruoti, Brent Roberts, and Kent Seamons. 2015. Authentication melee: A usability analysis of seven web authentication systems. 916–926.

[71] Phattarapong Sawangjai, Supanida Hompoonsup, Pitshaporn Leelaarporn, Supavit Kongwudhikunakorn, and Theerawit Wilaiprasitporn. 2019. Consumer grade EEG measuring sensors as research tools: A review. *IEEE Sensors Journal* (2019).

[72] Javad Sohankar, Koosha Sadeghi, Ayan Banerjee, and Sandeep K. S. Gupta. 2015. E-BIAS: A Pervasive EEG-Based Identification and Authentication System. 165–172.

[73] Nancy K. Squires, Kenneth C. Squires, and Steven A. Hillyard. 1975. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology* 38, 4 (1975), 387–401.

[74] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. 2019. Robust Performance Metrics for Authentication Systems.

[75] Shravani Sur and V. K. Sinha. 2009. Event-related potential: An overview. *Industrial Psychiatry Journal* 18, 1 (2009), 70.

[76] I. Svogor and T. Kisasondi. 2012. Two factor authentication using EEG augmented passwords. 373–378.

[77] Moin Syed and Sarah C. Nelson. 2015. Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood* 3, 6 (2015), 375–387.

[78] Kavitha P. Thomas, A. P. Vinod, et al. 2017. EEG-based biometrie authentication using self-referential visual stimuli. 3048–3053.

[79] Julie Thorpe, Paul C. van Oorschot, and Anil Somayaji. 2005. Pass-thoughts: Authenticating with our minds. 45–56.

[80] Marijn van Vliet, Christian Mühl, Boris Reuderink, and Mannes Poel. 2010. Guessing what's on your mind: Using the N400 in brain computer interfaces. In *Lecture Notes in Computer Science*. Vol. 6334 LNAI. Springer Berlin, Berlin, 180–191.

[81] Marijn van Vliet, Arne Robben, Nikolay Chumerin, Nikolay V. Manyakov, Adrien Combaz, and Marc M. Van Hulle. 2012. Designing a brain-computer interface controlled video-game using consumer grade EEG hardware. 6 pages.

[82] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2014. Honey, I Shrunk the Keys: Influences of Mobile Devices on Password Composition and Authentication Performance. 10 pages.

[83] Emanuel Von Zezschwitz, Anton Koslow, Alexander De Luca, and Heinrich Hussmann. 2013. Making graphic-based authentication secure against smudge attacks. 277–286.

[84] W3C. 2020. Web Authentication: An API for accessing Public Key Credentials Level 2. W3C Candidate Recommendation Snapshot. https://www.w3.org/TR/webauthn-2/.

[85] Frederick W. Wheeler, Richard L. Weiss, and Peter H. Tu. 2010. Face recognition at a distance system for surveillance applications. 8 pages.

[86] Jonathan Wolpaw and Elizabeth Winter Wolpaw. 2012. *Brain-Computer Interfaces: Principles and Practice*. Oxford University Press.

[87] Qunjian Wu, Ying Zeng, Chi Zhang, Li Tong, and Bin Yan. 2018. An EEG-based person authentication system with open-set capability combining eye blinking signals. *Sensors* 18, 2 (2018), 335.

[88] Mahendra Yadava, Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, and Debi Prosad Dogra. 2017. Analysis of EEG signals and its application to neuromarketing. *Multimedia Tools and Applications* 76, 18 (2017), 19087–19111.

[89] Su Yang and Farzin Deravi. 2017. On the usability of electroencephalographic signals for biometric recognition: A survey. *IEEE Transactions on Human-Machine Systems* 47, 6 (2017), 958–969.

[90] Hui-yen Yap, Yun-huoy Choo, and Wee-how Khoh. 2017. Overview of acquisition protocol in EEG based recognition system. In *Brain Informatics*, Zeng et al. (Ed.). Vol. 10654. Springer, Cham, Switzerland, 129–138.

[91] Xiang Zhang, Lina Yao, Chaoran Huang, Tao Gu, Zheng Yang, and Yunhao Liu. 2020. DeepKey: A multimodal biometric authentication system via deep decoding gaits and brainwaves. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 4 (2020), 1–24.

Received 8 February 2022; revised 4 November 2022; accepted 24 December 2022