IET Biometrics

The Institution of Engineering and Technology | WILEY

RESEARCH ARTICLE OPEN ACCESS

Stability of EEG-Biometrics: The Role of Artifacts in Within vs. Cross Session Authentication

Yvonne Höller[1] | Arne C. Bathke[2] | Andreas Uhl[2]

[1]Faculty of Psychology, University of Akureyri, Akureyri, Iceland | [2]Department of Artificial Intelligence and Human Interfaces, Paris Lodron University, Salzburg, Austria

Correspondence: Yvonne Höller (yvonne@unak.is)

Received: 8 September 2025 | **Revised:** 22 August 2026 | **Accepted:** 26 August 2026

Academic Editor: Pradeep Kanchan

Keywords: autoregressive model | biometrics | cognition | cross-validation | electroencephalography | pattern recognition

ABSTRACT

Artifacts in the electroencephalogram (EEG) have not been investigated systematically regarding their effect on EEG biometric performance. Since artifacts vary over time, they can be expected to lower EEG biometric performance in cross session cross-validation. However, because of their presumed uniqueness, they might increase within session cross-validation performance of an EEG biometric system. We examined the EEG biometric performance of 15 features within and across 6 sessions recorded over 3 days, including 3 different cognitive conditions, before and after artifact exclusion in a sample of up to 90 individuals. Our results revealed significant effects of artifact exclusion depending on the cross-validation scenario and the feature. Overall, within session cross-validation was not systematically affected differently by artifact exclusion as compared to cross session cross-validation, but depending on the feature and the type of cognitive task employed, artifact exclusion led to better or worse results, with some frequency-dependent features performing better when artifacts were excluded. Our results emphasize the overestimation of biometric performance in within session cross-validation (with equal error rates (EERs) down to 0) as compared to cross session cross-validation (with EERs for the same feature above 30) and the low reliability of results obtained when evaluating EEG biometric performance in small samples. We recommend that future developments of EEG biometrics must test their robustness against artifacts in the data.

1 | Introduction

Electroencephalographic (EEG) characteristics have been praised to be universal and impossible to circumvent or to forge [1]. Nevertheless, there exists doubt whether the EEG will ever be a useful biometric since there are imponderable obstacles when putting EEG biometrics into practice [2]. Specifically, variability of the EEG over time is a major challenge to achieve the stability in EEG biometrics [3–6]. Only a few studies tested the potential of the EEG as a biometric feature against its limitations due to the poor stability of EEG biometrics over time. The scientific gain of many studies published so far is limited because they did not account for EEG-variability over time, and the very same problem will prevent further progress in this field if scientists do not use longitudinal EEG data with at least two EEG recordings that are at least several hours, if not days or even months, apart. The size of this problem is evident by the accuracy of EEG biometrics reaching 100% in systems relying on one EEG recording but close-to-chance performance in cross-session validation with EEG recordings acquired on different days [7]. The brain activity captured with the EEG depends on current thoughts, amount of sleep in the night before the measurement, intake of any

substance that affects brain activity such as nicotine, caffeine, psychoactive medication, and also the time since the intake, affective state, mental state, and time of day in interaction with the chronotype of the individual. Furthermore, artifacts that appear in the data but do not stem from brain activity but from muscular or ocular activity, or from a variation in the quality of the contact between the electrode and the skin, play a role and are hard to control. Finally, virtually all medical conditions that are related to brain health, such as neurological and psychiatric diseases, affect reliability, with, for example, higher stability over time of epileptiform activity [8]. In view of all these factors affecting the replicability of any pattern in the EEG, it is of utmost importance that the performance of biometric systems based on EEG-features is evaluated with cross-session cross-validation [9, 10] and that the factors that affect the stability of EEG biometrics are investigated systematically. However, only a few publicly available data sets allow for benchmarking across several sessions. A recent data set [11] includes 95 individuals observed during two sessions, but the time interval between the two sessions varies across individuals. Probably due to a lack of adequate data, only limited investigation was conducted into the potential factors that affect stability over time. In the following, we review the approaches taken to analyze and establish the stability of EEG biometrics.

Session invariant features can be identified more reliably if more than two sessions are available: two sessions are used to extract session invariant features, and the third is used to test the model's performance. An early study with a sample size of $n = 9$ and three sessions on three different days [4] found that increasing the time between enrollment and authentication decreases biometric accuracy and claimed that training the system over multiple sessions recorded on three different days increased the performance. A more recent study developed a model based on a convolutional neural network (CNN) and achieved moderate performance in a small sample of 10 people with three sessions that were separated by 7–14 days [12]. Unfortunately, with these small sample sizes, the true performance can hardly be estimated, as it has been suggested that adding individuals to small samples decreases the accuracy of an EEG authentication system that seems to perform well on the smaller sample [13].

The second approach to boost the detection of invariant features of individual brain activity is therefore to acquire the EEG during specifically induced states, for example, obtained by visual stimulation [14–16], cognitive stimulation [17–20], or by a combination of multiple tasks [21, 22]. The idea behind this is to evoke a specific, reproducible brain activation, while recording the EEG at rest is prone to capturing the electric signature of mind-wandering thoughts a user experiences at the moment of acquisition. Early research in the field also suggested that an individual task for each person can enhance the uniqueness of the extracted features, as the individual task produces a mental password [4]. However, only a few studies that controlled the mental state implemented cross-session cross-validation, and those that did so were done on very small samples (Table 1 in [10]). The only studies with a larger sample of $n = 50$ found comparable performance of the system with and without cognitive stimulation [7, 26], and could not replicate the finding of an earlier study in $n = 9$ where performance was higher during a movement condition as compared to rest [5].

Nevertheless, it seems plausible that the choice of the cognitive stimulation also matters [22]. Therefore, the question remains of which kind of cognitive stimulation is best suited to generate the so-called mental password [4]. While the individual reaction to flashing lights might be similar across individuals, triggering memory provides access to a plethora of interindividual differences in semantic representations that could be measured as individual mental states [25]. As such, the cognitive task might not only decrease intraindividual variability over time by inducing a reproducible pattern of brain activation, but it might also increase interindividual variability, because cognitive performance relies on prior training and memories. It has been found that even within a single experiment, different stimuli can yield different equal error rates (EER) [26].

Another twist regarding cognitive stimulation is the lack of large-scale research using multisession data. It must be considered that cognitive task performance increases over multiple sessions, raising the question of how many sessions would be necessary to enroll a person in a biometric system based on EEG signatures obtained during cognitive activation. This is especially relevant for difficult cognitive tasks such as motor imagery, where individuals improve at eliciting consistent brain activity with repetition, that is, through training. Accordingly, it was found that EEG-biometric systems relying on motor imagery benefited from training the model over multiple sessions [4]. Most interestingly, another study examined imagined speech [28] and found that when training was performed in one session and testing data obtained from another session, the accuracy of the system was, on average, lower when the first session was combined with other sessions (average accuracy 88.73%) as compared to configurations where sessions two, three, and four (average accuracy 94.88%) were combined. Unfortunately, these studies were based on small sample sizes of $n = 9$ [4] and $n = 6$ [28]. A later study with $n = 50$ did not show a systematic advantage of combining session two with session three compared with combining session one with the other sessions [26]. Although it must be considered that the first recording session is not an ideal source of data for EEG biometrics, especially when cognitive stimulation is used, the usability of a system requiring multisession enrollment is not satisfactory.

More recently, EEG biometric features have been challenged towards more realistic settings by testing the interstate stability of features using different cognitive stimulation during enrollment and authentication [6, 30, 32]. This approach mimics the situation that even the same cognitive task might produce different patterns of brain activity in the very same individual. This is intuitively understandable if we consider a memory that is formed but will change slightly each time it is recalled, and it also changes with the experience of situations similar to that specific memory. Therefore, testing features against their performance when one cognitive task is used for enrollment and another for authentication mimics intraindividual variance in brain activity over time under controlled conditions, thereby allowing the experimental situation to study the effects of externally or internally induced cognitive states that vary across authentication sessions. As with cross session cross-validation, the correct recognition rate drops significantly from same-task to different-task scenarios [30, 32] unless the classifier is trained and tested on multiple states [32]. In this scenario, functional

TABLE 1 | Studies examining EEG biometric features in multiple sessions with cross session or cross-task cross-validation.

Study	*N*	No. of sessions	Time	Cross-validation	Artifact	Conditions	Features	Database
Marcel [4]	9	3	1 d	Cross session	Filtering	3	PSD 8–30 Hz	Own
Näpflin [23]	20	2	≥1 y	Cross session	Exclusion	1	PSD	Own
Kostilek [5]	9	2	1 y	Cross session	Filtering	3	AR 8–13 Hz	Own
Rocca [24]	9	2	1–3 w	Cross session	Filtering	1	AR	Own
Armstrong [25]	9–15	2–3	1 w–6 m	Cross session	None	1	ERP	Own
Das [26]	50	3	7–49 d	Cross session	None	1	Bandpower	Own
Maiorana [7]	50	3	7–34d	Cross session	Filtering	1	AR, PSD, COH	Own
Wang [27]	4	2	≥1 w	Cross session	ICA	1	Wavelet	Own
Ruiz [22]	20	2	48–516 d	Cross session	Filtering	4	ERPs	Own
Höller [10]	70	2	2 w	Cross session	Exclusion	1	AR/connectivity	Own
Brigham [28]	6	1	—	Cross-task	Exclusion	2	PSD from AR	Own
Yang [29]	108	1	—	Cross-task	None	6	Wavelet	Physionet
Fraschini [30]	108	1	—	Cross-task	ICA	6	CC, PLI, AEC, PLV	Physionet
Wang [31]	59/109	1	—	Cross-task	ICA	4	AR, PSD, En	Own/physionet
Wang [32]	59/109	1	—	Cross-task	ICA/MARA	4	AR/connectivity	Own/physionet
Wang [6]	109	1	—	Cross-task	Correction	4	CC, PSI	Physionet
Mota [33]	109	1	—	Cross-task	Filtering	4	DL	Physionet
Di [34]	10	3	≥2 w	Cross session	Filtering	1	PSD, PLI, coherence	Own
Rahman [35]	10	10	?	Cross session	VMD	1	PSD	Own
Ashenaei [36]	109	1	—	Cross-task	Filtering	4	Connectivity	Own/physionet
Plucinska [37]	29	20	>43 d	Cross session	Filtering	1	PSD	Own
Xu [38]	109	1	—	Cross-task	Filtering	4	DL	Physionet
Abdel-Ghaffar [39]	32/40	1	—	Cross-task	Correction	≥2	Various, DL	DEAP/SAM
Liu [40]	14/9	2	~23d	Cross session	Filtering	1/4	DL	Own
Fallahi [41]	345	>1	>1?	Cross-?	Filtering	?	AR, PSD, DL	PEERS
Usman [42]	21	1–3	?	Cross session	Correction	1	Various	BED

Abbreviations: AEC, amplitude envelop correlation; AR, autoregressive reflection components; CC, correlation coefficient; d, days; DL, deep learning; m, months; *N*, sample size; PLI, phase lag index; PLV, phase locking value; PSD, power spectral density; PSI, phase synchronization index; time, time between sessions; VMD, variational mode decomposition; w, weeks; y, years; ?, manuscript is not sufficiently clear on this item.

connectivity measures generalize better over diverse human states [31]. However, all studies employing cross-task cross-validation were performed for data collected in one session only (Table 1). We advocate that cross session/cross-task cross-validation is a more valid scenario.

The finding that functional connectivity measures are more stable across different cognitive tasks [31] leads to the third approach to finding session-invariant features, which focuses on the features themselves, with the hypothesis that some features are more stable than others. The event-related potential (commonly known as the ERP) is the most basic EEG feature. It was used in a study that combined several cognitive tasks with stunningly stable results even after years [22]. However, a protocol that requires several cognitive stimulations is highly time-consuming and cannot realistically be used as an authentication system. Furthermore, the signal quality of the ERP depends on the number of repetitions, so to achieve adequate biometric performance, a longer recording time is required to obtain sufficient repetitions, increasing the time needed to enroll and authenticate an individual. For example, in a prior study, 75 repetitions were required to obtain an ERP with satisfactory biometric properties [25].

In contrast, autoregressive components and features derived from the multivariate autoregressive (MVAR) model were found to be highly promising even with relatively short data [7, 10]. It was also proposed that the best feature vector of functional connectivity measures, as derived from MVAR coefficients, can be extracted using deep learning [31]. However, while this finding highlights the stability of the derived features across cognitive states, it was not tested across EEG sessions captured on different days. Moreover, the test–retest reliability of MVAR models and the features

derived from them is severely affected by artifacts [43], such that a very common and inexpensive solution is to use only artifact-free segments of the recorded signal. In research on EEG biometrics, artifacts were handled in many ways, including correction and even retention, with models designed to be invariant to added muscular artifacts [43]. One study claimed that an unpublished pilot work demonstrated that artifact rejection or filtering did not impact classifier accuracy when using the ERPs as a feature [25]. It was even suggested that artifacts can be fairly predictive and lead to high recognition rates because artifact patterns are more individually distinguishable than brain activity [28]. At worst, the excellent performance of an EEG biometric system might be due solely to individual user artifacts. Artifacts are an important factor that have been studied in EEG research in general, for example, with respect to various artifact removal techniques, but have largely been neglected in research on EEG biometrics and deserve systematic investigation for this specific application.

Table 1 lists prior studies with multiple sessions or multiple tasks and their handling of artifacts. To our best knowledge, there is no study that systematically compared the biometric performance of several EEG features across several sessions and different cognitive tasks, depending on the inclusion or exclusion of artifacts. Particularly, even very advanced studies comparing many datasets depending on whether they were collected in one or multiple sessions and that clearly state that the lack of multisession research is a shortcoming, do not adequately distinguish between multisessions collected on the same day or on different days [44], although the importance of this detail was emphasized clearly [35]. None of the prior studies implemented both cross-session and cross-task cross-validation. Also, for most studies, the tasks for cross-task cross-validation were drastically different, for example, rest vs. motor imagery or motor imagery vs. attention task, with a few exceptions where the same task with different stimuli was used [28]. We advocate that using the same task but with different stimuli better simulates the realistic scenario of changing brain responses over time to cognitive tasks due to learning, forgetting, and other unknown factors.

In summary, there is a need to investigate the following research questions:

- How does artifact exclusion affect biometric performance?
- How does biometric performance of different features benefit (or not) from the exclusion of artifacts?
- How does artifact exclusion affect biometric performance in different cognitive tasks?
- How does artifact exclusion affect cross session authentication performance?
- Do artifacts in the data affect how cross session validation performance is affected by the first vs. other sessions for enrollment/authentication?
- Do artifacts in the data affect how cross session validation performance varies as a function of days since enrollment?

The innovative contributions of this paper are represented by testing the following hypotheses:

1. Keeping artifacts in the data might augment uniqueness; however, it is possible that artifacts are subject to the current configuration of the EEG recording and underly other, currently unknown factors. We therefore hypothesize that artifacts might inflate within session performance but deteriorate cross-section performance.
2. We analyze EEG biometric performance across three different cognitive tasks, and each of them in 6 slightly different versions for the 6 sessions, each separated by at least 10 h, hypothesizing that biometric performance depends on the type of task used.
3. Since different features were found to vary in reliability over time depending on the inclusion or exclusion of artifacts [43], we expect that this is reflected in biometric performance, too.

2 | Methods

2.1 | Experimental Data

2.1.1 | Sample

The data collection of the sample used in this study was described previously [45–49]. These related publications provide details on the sample and recordings, but please note that the EEG data during cognitive stimulation have not been analyzed in any publication, so far. In the following, we summarize the most important details on data collection. The study and the analysis were not preregistered in an independent registry, but the study protocol was registered and approved by the ethical commission of the Region of Salzburg, Austria (Approval No. 415-E/1755/202016). The study was carried out in compliance with the Declaration of Helsinki, and written informed consent was obtained from all participants. We recruited and examined 106 people in the Epilepsy Monitoring Unit at the Department of Neurology, Paracelsus Medical University, Salzburg, Austria, between 2016 and 2018 within a larger study about the effect of epilepsy on memory. After excluding participants due to issues with cognitive testing and/or recording procedures, 90 participants remained for analysis. The first two participants were excluded because the cognitive testing procedure was changed thereafter. We excluded three patients who underwent invasive recordings instead of scalp recordings and four patients where the markers for the cognitive tasks were not registered in the EEG due to a connection failure in the hardware. Further participants dropped out because the cognitive recording session was missing. All participants underwent testing on the evening of the first night (6–8 pm) and then on each morning (8–10 am) and evening, up to 6 times in total. In each session they performed a verbal learning condition, which consisted of 60 presented word pairs, fingertapping (FT) of all individual fingers except for the thumb of the nondominant hand, and third, virtual reality (VR) recognition of scenes, which consisted in the presentation of 20 words describing objects to be remembered from a previous VR exploration of a town. Table 2 shows the total number of participants available for each cognitive task type and for the scenario before and after artifact exclusion.

The number of participants included for each cognitive task and for the artifact inclusion vs. exclusion scenario varies because we required a minimum of data available for each participant so that

TABLE 2 | Sample size in each cognitive task condition and session before (all) and after (AE) exclusion of artifacts.

Session	*n* all/AE FT	*n* all/AE WP	*n* all/AE VR
Session 1	86/85	90/74	68/70
Session 2	83/78	87/64	64/54
Session 3	75/69	81/52	62/40
Session 4	67/64	76/41	45/26
Session 5	60/53	58/26	32/16
Session 6	39/31	40/15	20/7

Abbreviations: AE, artifacts excluded; FT, Fingertapping; *n*, number; VR, virtual reality; WP, word-pair task.

the creation of enrollment templates and authentication experiments would be comparable between the two scenarios. For the same reason, the number of participants dropped over the course of the six sessions. The number of individuals for the artifact inclusion scenario for the VR task is lower than for the artifact excluded scenario because in two individuals the model estimation for the MVAR model did not converge due to low data quality.

2.1.2 | Data Registration

EEG was digitally recorded with the medical system SystemPlus Evolution SD LTM 46 Express Amplifier (Micromed S.p.A, Mogliano, Italy). Recordings began on the day of admission and lasted for up to 4 days (~96 h of continuous EEG) using 29 standard electrodes placed according to the 10–20 system. This setup of 29 electrodes was according to clinical routine requirements. The sampling rate was 1024 Hz and online filtering included a 0.1 Hz high-pass and 50 Hz notch filter (ground at Fpz and reference at Oz). Impedances were brought below 10 kΩ initially and monitored daily.

2.1.3 | Data Preparation

From the continuous 24 h recordings, we first segmented the relevant data according to markers that were placed during the 6 cognitive testing sessions. These segments were preprocessed with Brain Vision Analyzer (Version 2.1, Brain Products GmbH). Ocular correction according to the Gratton & Coles algorithm [50] was performed, using a vertical electrooculographic channel. Please note that although the ocular blinking pattern in the EEG represents an artifact that would be worth investigating within the research question of the present study, this artifact is so frequent in the EEG that it would lead to the exclusion of the majority of the data and, thus, making the artifact exclusion scenario unfeasible. Ocular artifacts stem from eye movements, usually blinking, and represent large deflections (hundreds of microvolts) of the signal beyond the variation of neurophysiological activity during wakefulness (tens of microvolts). This type of artifact distributes over the scalp with the fraction of the ocular signal decreasing with the distance of the scalp sensor from the eye.

2.1.4 | Artifact Removal

For the artifact exclusion scenario, an automatic artifact detection was carried out to exclude data contaminated by artifacts with the following thresholds as predefined in the well-established artifact removal in Brain Vision Analyzer: Maximal allowed voltage step per sampling point was 50 μV (values that exceeded this threshold were marked within a range of ±100 ms); maximal allowed absolute difference on an interval of 200 ms was 200 μV, and lowest allowed absolute difference during an interval of 100 ms was 0.5 μV (values that exceeded this were marked with a surrounding of ±500 ms). This method reliably detects muscular artifacts, sweat artifacts, movement artifacts, and other patterns that are not of neural origin.

2.1.5 | Data Segmentation

From the preprocessed data, we extracted single epochs for the three conditions. Continuous EEG data was segmented into epochs from 500 ms before stimulus onset to 1500 ms after stimulus onset for verbal learning and VR recognition and 50 ms before typing to 500 ms after typing for individual finger movements on the keyboard in the FT task. The preprocessed and segmented data was exported into a generic data format and imported to Matlab (release R2018b, The Mathworks, Massachusetts, USA).

2.2 | Feature Extraction

We used data from the 17 channels F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, O1, and O2. These channels represent the international 10–20 system and were chosen a priori without optimizing the selection of channels towards biometric performance. We would like to emphasize that this choice was solely based on the definition of these channels, representing the core minimum of electrodes in a typical low-resolution EEG setup. The other channels were excluded because these clinically required sensors are spatially closed to the selected electrodes and would mostly deliver redundant information. Thus, proceeding with a smaller set of electrodes reduced the computing time and allowed the feature optimization to run more efficiently. MVAR coefficients were estimated based on the MVAR model [51, 52], as given in Equation (1):

$$Y(t) = \sum_{k=1}^{P} A(k)\, Y(t-k) + U(t), \tag{1}$$

where the vector $Y(t) = [y_1(t),\ldots,y_M(t)]^T$ holds the values of the M sensors at time t, P represents the model order (in the present case 10), $A(k)$ are coefficient matrices of size $M \times M$ with elements $a_{ij}(k)$ describing the dependence of $y_i(t)$ on $y_j(t-k)$. Finally, $U(n)$ is the innovation process segments of the signal of length n. We relied on the BioSig toolbox [53] and used the functions mvfreqz.m and mvar.m to fit the model using partial correlation estimation with unbiased covariance estimates [51], which was recommended previously [54]. The model order was chosen to be $P = 10$ to keep our analysis comparable to prior work [7]. The matrices $A(k)$ represent the multivariate autoregressive coefficients (AR), that is, the main feature analyzed in the present study.

To illustrate the problem's range across features, we additionally extracted a set of features commonly used in the literature to keep our analysis comparable to previous work (Table 1). Please refer to

a previous study where we describe these features in more detail [10], although their extraction in that reference was based on slightly different parameters (i.e., a higher model order and a different dataset). The single-sided amplitude spectrum from the Fast Fourier Transform of the signal served as the power spectral density (PSD), that is, another feature analyzed in the present study. The PSD was obtained with the Matlab function fft after removing constant mean trends from each segment's time course. Another set of features can be extracted from the estimated MVAR model by transforming it from the time domain into the z-domain and the f-domain, yielding two transfer functions. The other multivariate parameters were derived from these transfer functions for 1 Hz frequency steps between 1 and 48 Hz (literature indicating the details on how to derive these features is referenced alongside with each feature but please see Ref. [10] for further explanations on the individual features): spectrum (S) [55], transfer function (hh) and transfer function polynomial (AF) [56], coherence (COH) [57], partial coherence (pCOH) [58], partial directed coherence (PDC) and partial directed coherence factor (PDCF), generalized partial directed coherence (GPDC) [59], directed transfer function (DTF) [60], direct directed transfer function (dDTF), full frequency directed transfer function (ffDTF) [61], and Geweke's Granger Causality (GGC) [62, 63]. Additionally, direct causality (DC) [64] was obtained as a frequency-independent measure. PSD and all frequency-dependent measures of interaction were averaged in classical frequency ranges delta (1–4 Hz), theta (5–7 Hz), lower alpha (8–10 Hz), higher alpha (11–13 Hz), lower beta (14–20 Hz), higher beta (21–30 Hz), and gamma (31–48 Hz).

2.3 | Feature Vectors

Depending on whether the feature was directed or not, the obtained interaction matrices were symmetric (not-directed measures, e.g., COH) and therefore contained redundant elements, or not symmetric (directed measures, e.g., DTF). The feature vectors were obtained by concatenating all nonredundant values. Therefore, the feature vectors were of length $17 \times 17 = 289$ (DC without frequency dimension) up to $17 \times 17 \times 7 = 2023$ for directed measures.

2.4 | Templates for Enrollment and Authentication

Enrollment in an authentication system requires registering a biometric data template that can be matched to the user during authentication. For authentication, previously unseen data is compared to that template. In our study, template creation for enrollment and authentication follows the same process. A template was formed by the average of the feature vectors v of several epochs $e_1 \ldots e_k$ that were recorded from one participant p. Thus, one element i in the template feature vector V for a template of participant p can be described as follows in Equation (2):

$$V_p(i) = \frac{1}{k} \sum_{l=1}^{k} v_{p,e_l}(i). \tag{2}$$

Each template could be a template for enrollment or authentication. We averaged over $k = 8$ epochs for the verbal condition, 3 epochs for the VR condition, and 12 for the FT condition in order to obtain 5 templates for each condition and each participant p. Note that these epochs do not necessarily need to be all of the epochs that belong to one participant, but these numbers were a compromise to use all data available for the participants with the lowest amount of data available. Therefore, for participants with more data, the redundant epochs were discarded from the end of the data stream. Creation of a template for the enrollment or authentication process for the verbal condition required recording 16 s of EEG, 6 s for the VR condition, and 6.6 s for the FT condition. For the scenario when artifacts were excluded, data for individual sessions and participants were used if there was a minimum number of epochs left for the creation of the templates after exclusion of epochs overlapping with artifacts. The minimum was 40 epochs for the verbal learning condition (out of a total of 60), 15 epochs for the VR condition (out of a total of 20), and 60 epochs for the FT condition (out of a variable total number, depending on typing speed).

2.5 | Comparison Scores for Enrollment and Authentication

To verify whether a person is in the biometric database, the biometric data the person provided during the authentication process must be compared to the templates stored in the database. To this end, a comparison score needs to be computed between the person's template that serves as the authentication probe and all templates in the database. In the present study, the comparison score Z in Equation (3) was the Pearson correlation ρ of two templates V from person p_a and person p_b.

$$Z = \rho\left(V_{p_a}, V_{p_b}\right). \tag{3}$$

Templates V_{p_a} and V_{p_b} match the same person if $a = b$. The Pearson correlation coefficient between feature vectors as an index of similarity between two templates allows for adequate scaling across varying feature vector length.

2.6 | Computation of Performance Metrics

Mated scores are obtained as Pearson correlation coefficients between all 5 templates of one individual. Accordingly, the nonmated scores were obtained for each individual by calculating the Pearson correlation coefficients between each individual's templates and all templates from all other individuals. All mated scores (i.e., template correlations within subjects) and nonmated scores (i.e., template correlations between subjects) were collected in two separate distributions from which we determined the equal error rate (EER) as the intersection of the two empirical distributions, that is, the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR).

2.7 | Feature Selection Algorithm

The feature vectors likely contained redundant information due to similar values across adjacent channels or frequency ranges. Therefore, a feature subset selection algorithm was implemented in a three-layered cross-validation procedure:

1. In the outer layer we implemented 5-fold cross-validation. We left 1/5 of the participants out in each iteration as the *test set*, while the other 4/5 formed the set for feature selection and training, submitted to the middle layer. This partition was random, but to ensure comparability across

features, we used the same partition into 5 sets for all features, all conditions, sessions, and session combinations for cross-session cross-validation.

2. In the middle layer, these 4/5 of the sample were again divided for a 5-fold cross-validation. In each iteration, 1/5 of the participants were left out, the other 4/5 were submitted to the inner layer as the training set. Also here, a random partition was created that was used for all features, conditions, sessions, and session combinations.

3. In the inner layer, the *training set* was used to sort and select features:
   - For each entry in the feature vector, we divided the mean of all participants' p standard deviation across all epochs σ_e—which should be small for a good feature—by the standard deviation over all participants average epoch σ_p (i.e., the standard deviation across all templates V of all subjects p), which should be large for a good feature as described in Equation (4):

$$\sigma_\delta(i) = \frac{\overline{\sigma_e(v_p, i)}}{\sigma_p\,(V(i))}. \tag{4}$$

   - We sorted the entries $i = 1\ldots k$ of the feature vector by the ratio σ_δ from smallest to largest. This approach efficiently shortens the feature selection procedure and is a viable alternative [10, 65] to principal component analysis [66, 67] or independent component analysis (ICA) [68].
   - The first 4 entries of the sorted feature vector formed the initial subset of selected features.
   - We calculated the EER for this subset.
   - Then, we iterated over the feature vector by adding one entry at a time to the subset of selected features and calculating the EER for this new feature subset. If the resulting EER was smaller as compared to the last EER, the added feature was kept in the selected feature subset, but it was excluded otherwise. When 50 consecutively tested entries were tested but not kept, we stopped the procedure and did not iterate over the rest of the feature vector.

4. The selected features formed the optimized feature vector. Since this feature subset selection was done 5 times, this resulted in 5 optimized feature vectors. The average length of the 5 feature vectors was calculated, and the features from the 5 vectors were sorted according to how often they were selected in the 5 iterations. The final feature vector included the top-most selected features and had the length equal to the average length of the 5 optimized feature vectors.

5. The final feature vector was then used to calculate the EER based on the initially left-out *test set* from the outer layer. However, since the mated and nonmated scores were based on 5 different lists of feature vectors, we once again ranked the features by how often they were chosen in those 5 outer-layer loops and selected the most frequently chosen features from this list. The final number of the topmost selected features was the average length of the 5 feature vectors across the 5 cross-validation iterations. From this final list, we calculated the final EER that was used to compare the different scenarios.

Figure 1 illustrates the cross-validation procedure.

2.8 | Statistical Comparisons

We tested…

1. … the effect of artifact removal for all 3 conditions, 15 features, and 21 cross-validation scenarios by comparing the EERs with vs. without artifact removal.
2. … differences between the EERs obtained in within-session and cross-session cross-validation by comparing session 1 to the other 5 within-session scenarios as well as to the cross-session scenarios 12, 1–3, 1–4, 1–5, and 1–6. We limited this analysis to session 1 vs. all others because that allowed us to substantiate the analysis with the largest sample sizes.
3. … the effect of an increasing time between enrollment and authentication on EERs by comparing cross session cross-validation from session 1–2 vs. 1–3, 1–4, 1–5, and 1–6.

In order to compare two EERs we designed the following randomization test procedure: We merged the matching scores of

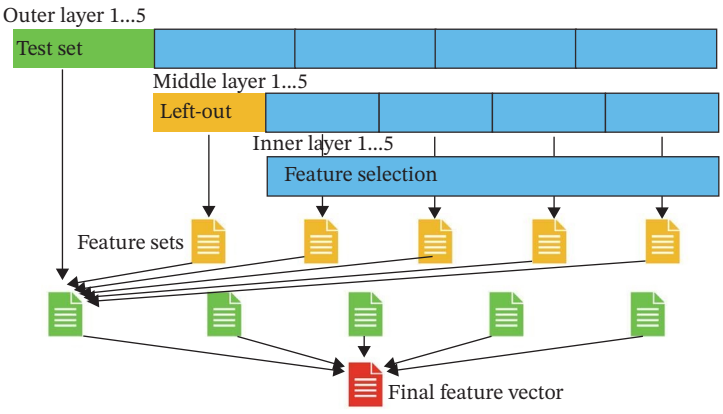

FIGURE 1 | Procedure of cross-validation and feature selection. Test set performance metrics are calculated from the five outer-layer feature sets (green). Final performance metrics are calculated from the final feature vector (red).

both conditions, A with $N = 1 \ldots n_a$ matching scores and condition B with $N = 1 \ldots n_b$ matching scores, and randomly drew two sets of matching scores of size n_a and n_b that would include matching scores from both conditions A and B. We performed the same procedure for the impostor scores. With these 4 sets of random matching and impostor scores, we calculated again the two random EERs, that is, $EER_R{}^A$ and $EER_R{}^B$. We calculated the difference between these two EERs, that is, $EER_R{}^\Delta$ under random conditions. We repeated this procedure a $P = 1000$ times. Then, we compared the actually observed difference $EER_{obs}{}^\Delta = EER_A - EER_B$ to the distribution of 1000 random EER differences $EER_R{}^\Delta$. The proportion of random EER differences that was more extreme, that is, larger in its absolute value than the original EER difference, will from now on be denoted as the p-value of the test (Equation (5)). Although strictly speaking, it does not fulfill all the properties of p-values in statistical hypothesis testing, it may be interpreted in a somewhat analogous way.

$$p = P\left(\left|EER_R^\Delta\right| \geq \left|EER_{obs}^\Delta\right| \middle| H_0 \right). \quad (5)$$

Because the different within session and cross-session scenarios include a varying number of individuals, we also conducted an additional test for the purpose of conducting an overall comparison of the artifact inclusion vs. exclusion scenario, separately for the within- and cross-session cross-validation scenarios. To this end, we merged all matching scores across all within-session scenarios and all nonmatching scores across all within-session scenarios and calculated an overall EER for the within-session scenarios, but separately for each feature. We repeated the same procedure for the cross session scenarios. We conducted the same statistical comparison of artifact inclusion vs. exclusion as described above for these EERs.

The resulting p-values were corrected with Bonferroni–Holm correction by merging all p-values of all tests conducted. According to this procedure, only test results with a p-value $p < 0.00004$ were declared to be significant.

3 | Results

Supporting Information Figures S1–S15 in the supporting section show for all 15 features the sub-selection of feature vector entries that were most commonly chosen across all cross-validation scenarios and tasks. For several features, autocorrelation of channels was more likely to be selected in the optimization procedure (features Af, DC, DTF, ffDTF, GPDC, hh, pCOH, PDC, PDCF, and S) and channels in the occipital and parietal lobe (features Af, AR, COH, DC, GGC, hh, pCOH, PDC, PDCF, PSD, and S) For the AR feature, lower indices indicating a shorter time gap in the regression model were more likely chosen, indicating the relative importance of faster repetitive patterns in the EEG. This is also consistent with the frequency-dependent features, which showed a higher likelihood of features to be chosen in the frequency range of 14 Hz and higher, and especially 31–48 Hz (features COH, dDTF, ffDTF, hh, pCOH, PSD, and S) with some exceptions where slower frequencies where more important (feature GGC, PDC, and to some extent PDCF).

Table 3 shows the average and minimum EERs across features for the different cognitive tasks, cross-validation scenarios, and artifact exclusion scenarios. For all three tasks, within-session cross-validation led to lower EERs than cross-session cross-validation. For all tasks and cross-validation scenarios, excluding artifacts led to lower EERs than keeping them in the data. When looking only at the reliable results from cross-session cross-validation, the best results were obtained in the word-pair task, followed by the VR task and finally the FT task.

TABLE 3 | EER averages/minima across features and sessions.

		Average EER		Minimum EER	
Task	**Cross-validation**	**AE**	**AI**	**AE**	**AI**
Finger tapping	Within	9.42	20.06	0.19	7.79
	Cross	24.61	26.86	9.05	12.84
Virtual reality	Within	14.49	14.43	0.27	1.96
	Cross	26.27	29.55	4.72	13.58
Word-pair	Within	6.24	10.24	0.01	1.26
	Cross	21.17	24.23	2.95	8.34

Abbreviations: AE, artifacts excluded; AI, artifacts included; cross, cross session cross-validation; EER, equal error ratio; Within, within session cross-validation.

Table 4 shows the average and minimum EERs for cross-session cross validation only, each feature individually, averaged across the tasks and across the 15 combinations of sessions. On average, most features benefit from artifact exclusion, resulting in a lower

TABLE 4 | EER averages/minima across tasks and sessions for cross-session cross validation.

	Average EER		Minimum EER	
Feature	**AE**	**AI**	**AE**	**AI**
S	25.22	33.28	5.89	22.68
DC	30.94	28.32	18.65	22.21
AR	30.17	23.57	19.91	15.83
hh	16.76	23.72	2.95	14.20
PDC	26.14	27.60	15.50	17.99
COH	19.67	25.73	6.72	13.80
DTF	22.86	30.14	11.47	19.01
dDTF	20.83	22.26	5.10	12.60
ffDTF	18.08	23.93	5.35	12.84
pCOH	17.36	18.94	5.08	8.34
PDCF	30.90	30.99	20.37	22.00
Af	21.07	22.03	9.89	14.40
GPDC	26.63	29.44	18.09	21.06
PSD	22.58	30.29	6.72	23.85
GGC	31.01	32.94	21.06	22.74

Abbreviations: AE, artifacts excluded; AI, artifacts included; EER, equal error ratio.

EER. However, for the features, DC and AR show lower EER averages and minima when artifacts are kept in the data. Lowest EERs under 10 are found after artifact exclusion for the features S, hh, COH, dDTF, ffDTF, pCOH, Af, and PSD. pCOH is the only feature resulting in minimum EERs below 10 both with and without artifacts in the data.

Figure 2 gives an overview over the EERs obtained for each cross-validation scenario (on the *y*-axis), all features (on the *x*-axis), for the three cognitive tasks (three rows of plots), and depending on whether artifacts were excluded or not (columns of plots). Before addressing the research questions individually in separate subsections of the results, it is of interest to point out several general observations. Firstly, a core observation from Figure 2 is that, as expected, within-session cross-validation (the first 6 rows in all plots) yielded, in general, lower EERs than cross-session cross-validation with perfect EERs = 0, for example, for features COH and PCOH, for most sessions 1–6. EERs were considerably higher in the cross-session cross-validation with EERs >30 for most session-combinations and features. Secondly, the figure gives

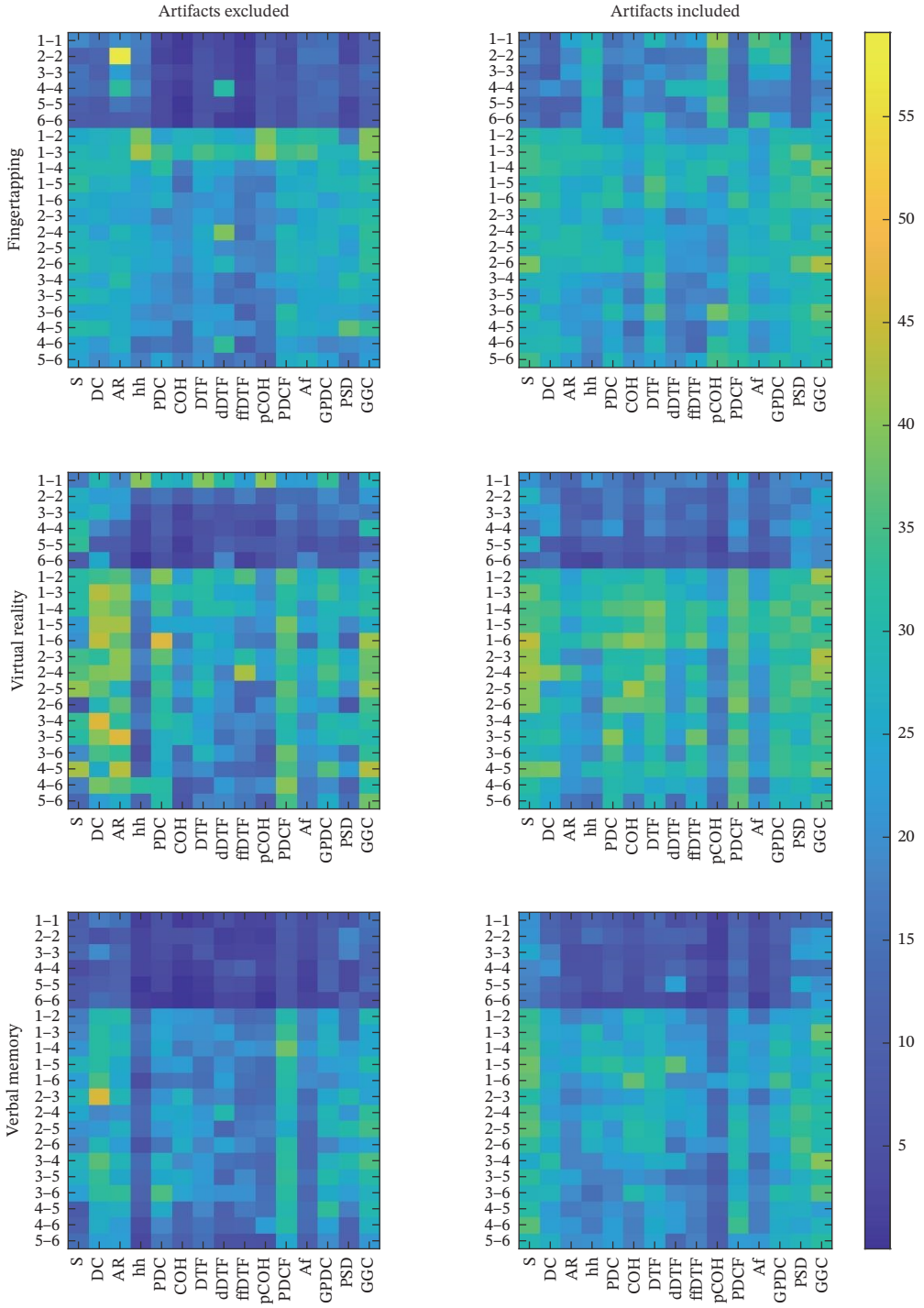

FIGURE 2 | Color-coded EERs for cross-validation scenarios on the *y*-axis, features on the *x*-axis, the three cognitive tasks fingertapping, virtual reality, and verbal memory in the three rows of plots, and depending on whether artifacts were excluded (left column) or not (right column of plots).

the impression that artifact exclusion might yield lower EERs for most features, especially for within session cross-validation.

3.1 | How Does Artifact Exclusion vs. Inclusion Affect Biometric Performance?

According to Tables 3 and 4, on average, artifact exclusion benefits biometric performance for most features. However, to answer the research question of how artifact exclusion affects biometric performance, we statistically compared the EERs obtained from data after artifact exclusion with the EER obtained from data where artifacts were included. Figure 3 shows the result of this comparison separately for features (columns), cognitive tasks (rows), and within session (top plot), as well as cross-session (bottom plot) cross-validation, but merged across the 6 sessions. Most features and tasks benefited more from excluding artifacts, both in the single session and cross-session cross-validation scenario (dark blue).

3.2 | How Does Biometric Performance of Different Features Benefit (or Not) From the Exclusion of Artifacts?

Regarding our second research question on how the biometric performance of different features benefits (or not) from the exclusion of artifacts there were several differences between features that we could note (Figure 3). Keeping artifacts in the data was advantageous across all three tasks and both cross-validation scenarios for the feature DC only. The features ffDTF, Af, and PSD consistently benefitted from excluding artifacts for all tasks and both within- and cross-session cross-validation scenarios. Notably, the features of ffDTF and PSD are frequency-dependent features, while DC is a frequency-independent feature.

3.3 | How Does Artifact Exclusion Affect Biometric Performance in Different Cognitive Tasks?

The EER in the single session cross-validation for the FT task was lower after exclusion of artifacts for most features (Figure 3). The EER obtained from the verbal memory task showed a similar pattern, but the VR task showed lower EER when artifacts were included for several features, especially in the single session cross-validation scenario.

3.4 | How Does Artifact Exclusion Affect Cross Session Authentication Performance?

Figure 4 shows the significance test results for comparing EERs in the artifact inclusion vs. exclusion scenario for the three cognitive tasks and all cross-validation scenarios individually. As is evident already from Figure 3, cross-session cross-validation did not systematically differ in terms of the advantage or disadvantage of excluding artifacts. Rather, the artifact exclusion affected the performance of different features differently. The only exception to this trend was the cross-session cross-validation scenario of the 5th vs. 6th session, which consistently yielded significantly

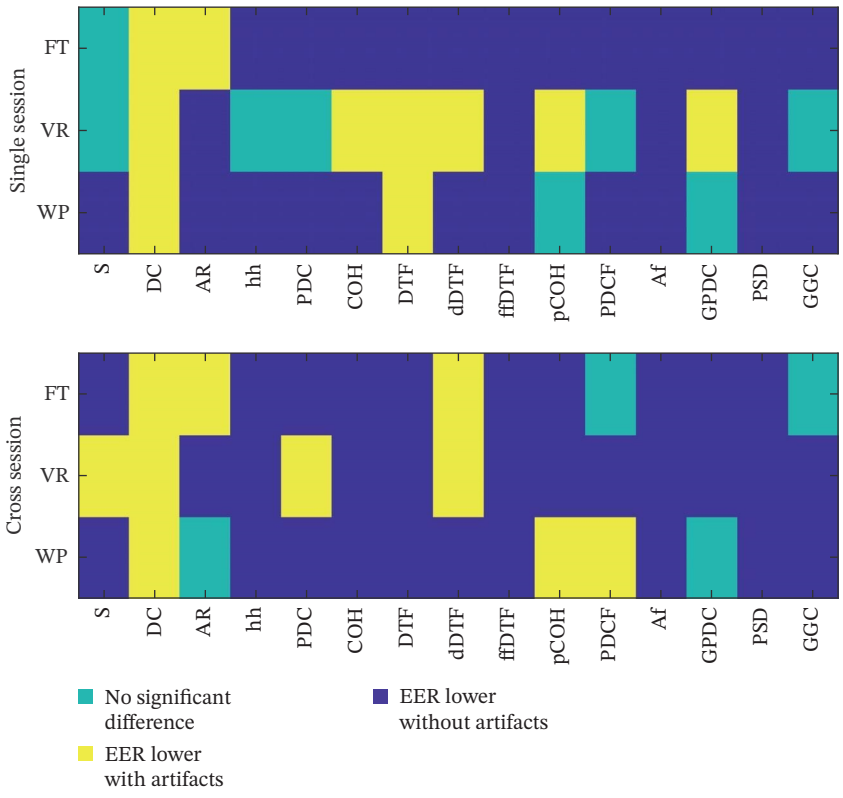

FIGURE 3 | Color-coded significance of statistical comparison of EERs of artifact exclusion vs. artifact inclusion for all within-session cross-validation scenarios based on single sessions (top) and all cross-session cross-validation scenarios based on pairs of sessions (bottom). The three cognitive tasks fingertapping (FT), virtual reality (VR), and verbal memory (WP) are given as separate rows on the *y*-axis and features on the *x*-axis. Dark blue indicates significantly lower EER when artifacts were excluded from the data, yellow indicates significantly lower EER when artifacts were kept in the data.

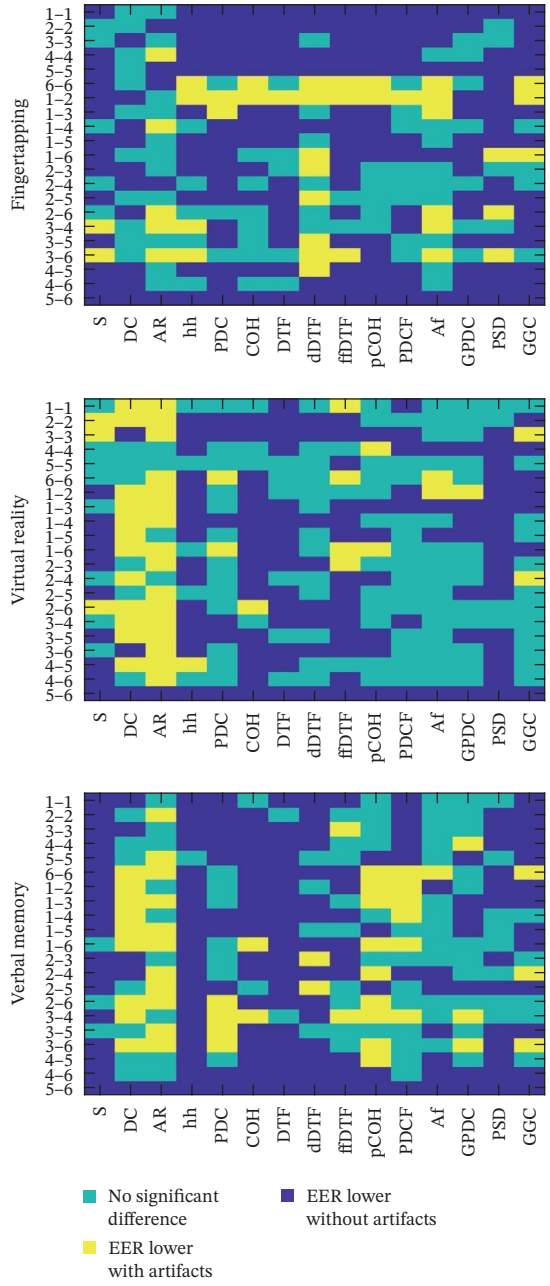

FIGURE 4 | Color-coded significance of statistical comparison of EERs obtained from artifacts exclusion vs. no exclusion data for all cross-validation scenarios on the *y*-axis, features on the *x*-axis, the three cognitive tasks fingertapping, virtual reality, and verbal memory in the three rows of plots. Dark blue indicates significantly lower EER for artifact exclusion, yellow indicates significantly lower EER for the artifact inclusion.

lower EER for artifact exclusion data across all features and tasks. This result could most plausibly be explained by the very small subsample of participants in this scenario.

3.5 | Do Artifacts in the Data Affect How Cross Session Cross-Validation Performance Is Affected by the First vs. Other Sessions for Enrollment/Authentication?

Figure 5 shows the results of comparing biometric performance from within session cross-validation based on session 1 to all other cross-validation scenarios. Within session cross-validation yielded lower EERs than cross-validation scenarios regardless of whether artifacts were kept in the data or excluded for the verbal memory task and for the FT task after exclusion of artifacts, as well as the VR task when including artifacts into the data. We expected no significant differences between within session cross-validation scenarios but found that they did differ significantly in many instances.

3.6 | Do Artifacts in the Data Affect How Cross-Session Cross-Validation Performance Varies as a Function of Days Since Enrollment?

According to the statistical results shown in Figure 6 there was no overall pattern suggesting that the EER deteriorates with increasing time since enrollment (which would be indicated by dark blue in the figure). Especially when artifacts were excluded, there were several features, which seemed to show the opposite pattern, that EER was lower in cross-task session combinations with longer time in between sessions. When artifacts were included, for some features, the shorter time between enrollment and authentication was beneficial. However, the pattern was not consistent across all features and tasks.

4 | Discussion

In the present research, we aimed to examine how artifact exclusion from EEG data affects biometric performance. In the following, we answer our research questions, based on the results presented in the previous sections.

4.1 | How Does Artifact Exclusion Affect Biometric Performance?

As a main conclusion, whether to exclude artifacts depends mainly on the feature and on other factors, such as the cognitive task. However, by tendency, EERs are lower after exclusion of artifacts from the data.

4.2 | How Does the Biometric Performance of Different Features Benefit (or Not) From the Exclusion of Artifacts?

Our data suggested that EERS obtained from some frequency-dependent features performed better for data after artifact exclusion, while for some frequency-independent features, the artifacts seemed to provide additional information. However, it should be noted that the advantage of excluding artifacts could not be observed clearly across all tasks, and this pattern was not replicable across all cross-validation scenarios. While the coefficients represented in the feature AR are a popular feature in EEG biometrics, our research shows that its sensitivity to artifacts in the data within- and cross-session cross-validation scenarios and different tasks deserves more attention. Notably, all other features presented here, except for the PSD, are derived from the AR, so it seems to be, to some extent, possible to circumvent this sensitivity to artifacts with further processing of these coefficients.

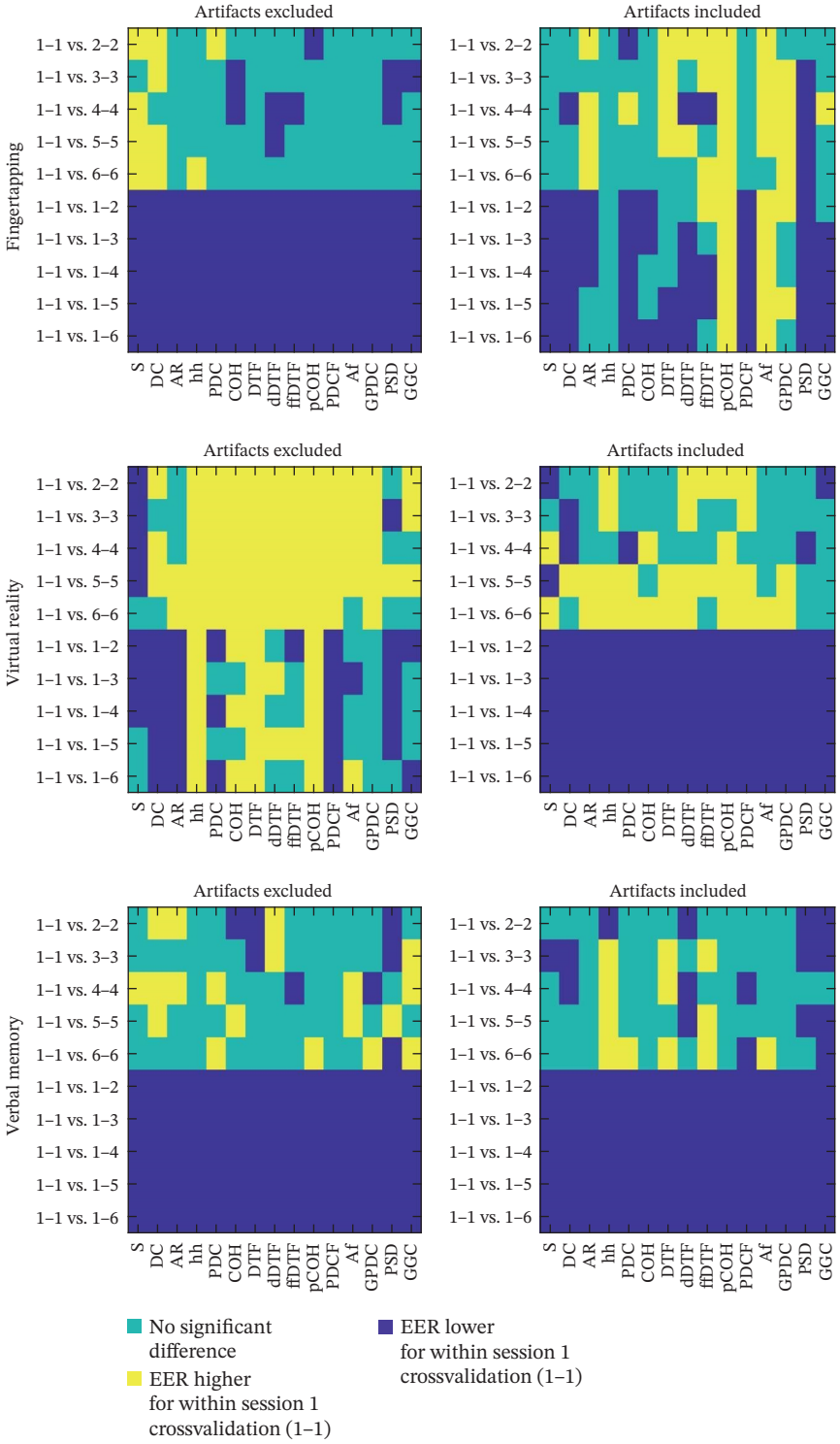

FIGURE 5 | Color-coded significance of statistical comparison of EERs of within-session cross-validation for session 1 vs. all other sessions and session 1 vs. other session cross-validation scenarios obtained from artifacts exclusion (left column) vs. no exclusion (right column) data. Comparison of cross-validation scenarios is represented on the *y*-axis, features on the *x*-axis, the three cognitive tasks fingertapping, virtual reality, and verbal memory in the three rows of plots. Dark blue indicates significantly lower EER for within-session 1 cross-validation, yellow indicates significantly lower EER for the other cross-validation scenario.

4.3 | How Does Artifact Exclusion Affect Biometric Performance in Different Cognitive Tasks?

We compared three cognitive tasks. While the comparison is limited due to a varying amount of data available for each task, it allowed to assess whether the effect of artifact inclusion or exclusion would be consistent across these different scenarios. Only a minority of features showed consistent results across the three cognitive tasks regarding the effect of artifact exclusion. It must be considered that the FT task differed from the other two tasks (a) in terms of the brain regions involved, (b) the available

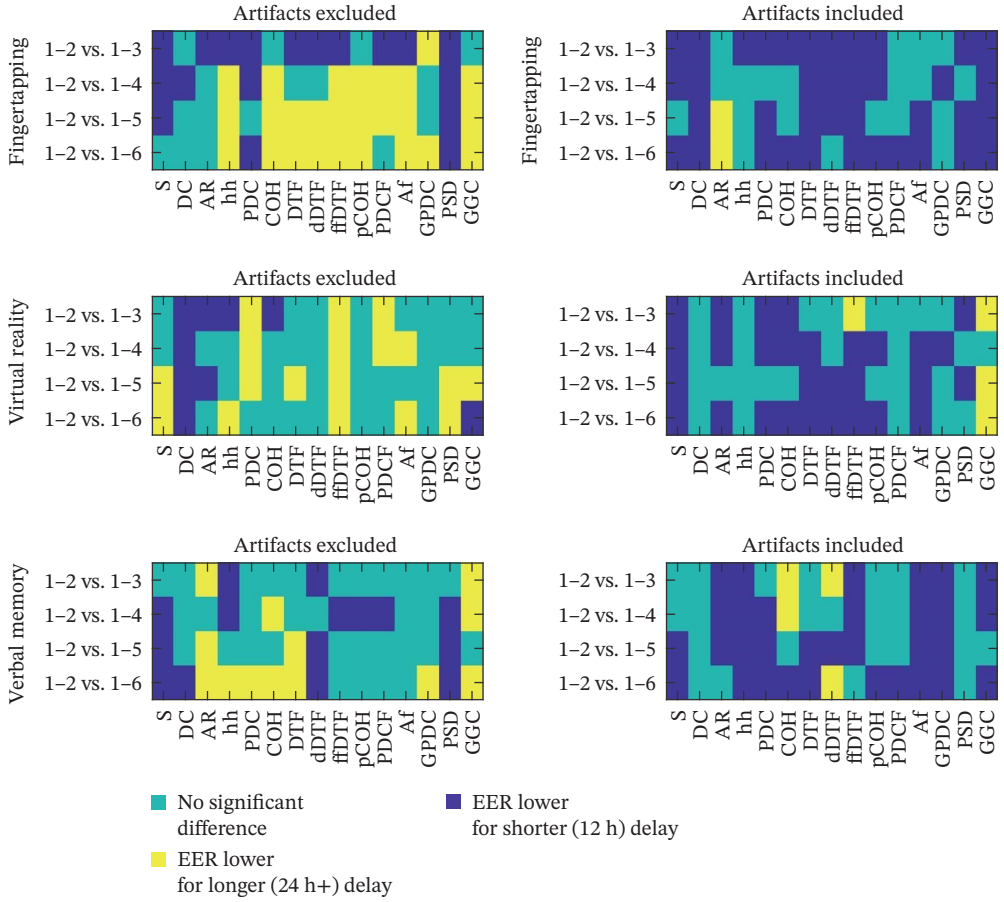

FIGURE 6 | Color-coded significance of statistical comparison of EERs of cross-session cross-validation for sessions 1 and 2 vs. 1 and all other sessions obtained from artifacts exclusion (left column) vs. no exclusion (right column) data. Comparison of cross-validation scenarios is represented on the *y*-axis, features on the *x*-axis, and the three cognitive tasks, fingertapping, virtual reality, and verbal memory, in the three rows of plots. Dark blue indicates significantly lower EER for cross-session cross-validation with enrollment in session 1 and authentication in session 2, yellow indicates significantly lower EER for the other cross-validation scenario.

sample size, and (c) the short segments extracted from the data due to the nature of this task. The length of epochs for analysis likely affects the calculation of the MVAR model because of a larger number of samples available for model estimation. Also, the FT task aimed at movement-related EEG correlates where the most prominent activity changes are confined to the motor cortex, covered mostly by 2 sensors, while the other two tasks recruited neural networks of episodic (VR task) and semantic (verbal memory task) memory that rely on spatially widespread neocortical memory storage. Since biometric performance in scenarios with different cognitive conditions during enrollment and authentication depends on channel density [6] it is possible that the effect of artifacts also changes as a function of sensor density in interaction with cognitive tasks. Furthermore, cognitive conditions had a varying number of epochs in the average that formed the template, with more epochs in FT and verbal memory (12 and 8) and fewer in the VR task (3 epochs). We would have expected that a larger number of epochs contributing to the template average, as it is the case for the FT and wordpair task, would make the results more resilient against artifacts. However, we did not note a larger difference between artifact inclusion vs. exclusion for the VR task as compared to the verbal memory task. The number of epochs used to create a template, thus, seems to play a minor role with respect to artifacts in the data.

4.4 | How Does Artifact Exclusion Affect Cross-Session Authentication Performance?

Since within-session cross-validation is not a valid way to assess the performance of a biometric system, emphasis should be given to the effect of artifact exclusion on cross-session cross-validation. The low validity of within-session cross-validation is further underscored by the variation in results when comparing different sessions, tasks, and features. We would have expected to find no difference among the various within-session cross-validation scenarios regarding whether artifacts are excluded, but considerable differences emerged, which might be explained by differing sample sizes and sample compositions across these scenarios (as mentioned in [13]). Although it is not obvious why the EER should be sample-size-dependent, any associated "*p*-value" or "significance" would certainly be expected to vary across sample sizes and compositions. Our results also emphasize that within-session cross-validation should not be used to benchmark an EEG biometric system and that the performance of EEG biometrics should not be evaluated on small samples. However, artifact exclusion did not consistently improve cross-session cross-validation results. The effect of artifact exclusion depended more on the feature than on the cross-validation scenario.

4.5 | Do Artifacts in the Data Affect How Cross-Session Validation Performance Is Affected by the First vs. Other Sessions for Enrollment/Authentication?

Most features yielded varying patterns over sessions and cognitive task conditions. For most features, excluding artifacts did not make different sessions more comparable to each other. Only for the FT task, excluding artifacts led to slightly more similar results across features in the within session cross-validation.

4.6 | Do Artifacts in the Data Affect How Cross Session Validation Performance Varies as a Function of Days Since Enrollment?

There was no clear pattern that would suggest that artifacts in the data affect how cross-validation performance varies as a function of days since enrollment. More generally, we did not see a clear decline of biometric performance as a function of days, whether it was based on data including or excluding artifacts. By tendency, keeping artifacts in the data led to better results when the time-span between enrollment and authentication was shorter for some features and some conditions.

4.7 | Future Directions

The tests in the present work were limited to handcrafted features, but deep learning-based algorithms are considered highly effective for EEG–based biometrics [12, 43]. Deep-learning-based algorithms are powerful, but it still needs to be demonstrated how deep learning models are robust against artifacts, and the excellent performance of deep-learning algorithms needs to be verified in larger databases of longitudinal EEG data. In addition to varying state of mind and artifacts, it was also claimed that the electrode positions might change from recording to recording, thus making it difficult to handle several sessions as comparable data sets [5]. In the present work, electrodes were glued onto the head on the day of the first recording and, therefore, were the same up to the last recording session. While we were able to exclude the effect of slight variations in electrode positions, it is worth considering a systematic investigation of the variation of electrode positions between sessions with realistic distances in the millimeter range.

Our results could also be interpreted as suggesting that the time of day might play a role for some features. While sessions 1, 3, and 5 took place in the evening, sessions 2, 4, and 6 took place in the evening. So far, only a few studies investigating cross-session stability of EEG biometrics controlled the time of the day, with all recordings being performed between 9 am and 12 am [23] or with both sessions taking place at the same time of the day [10]. While we found no overall systematic effect of time of day in relation to the artifacts, some features showed an on–off pattern in the significance of test results. We suggest that it might be worth looking into the effect of time of the day in the sense of enrollment at one time, for example, the morning, and authentication at another time, for example, the evening. This effect should be examined alongside with information on the chronotype of the individuals, which is, to our best knowledge, not available in the open databases that have been used for EEG biometrics, so far.

Progress in the field is also hampered by a lack of consistency in terminology. In our research, we refer to cross-session cross-validation if the two EEG recordings are separated by several hours, in the present work, at least 10 h when one recording took place in the morning and the other one in the evening. However, recent work did consider different runs of an experiment as different sessions [36].

The EEG can reveal sensitive information, such as medication [69]. Therefore, privacy is an issue [70]. This is especially true for the present data set, which was acquired in a clinic and, therefore, cannot be shared. Because of privacy concerns, cryptographic components were introduced into biometric systems that rely on EEG [71]. This sensitive nature of the signal prevents quick progress in research, as large open-access EEG data sets that allow for testing biometric systems under realistic circumstances are not easily available to allow benchmarking of technical achievements [11]. Additionally, the time and effort required to record an EEG contribute to high dropout rates in studies aiming to obtain multiple EEG recordings over a longer period [22, 25]. Future endeavors to establish large databases, including data from neurologically healthy individuals as well as neurodiverse participants with at least two sessions will help the field to develop inclusive and robust EEG biomarkers.

4.8 | Limitations

While we systematically investigated the effect of artifacts in the data, it was not possible to keep all other factors constant across the scenarios we examined. Most importantly, we observed differences between scenarios with larger vs. smaller sample sizes, which is most likely explained by the sample size. We could have based the whole analysis on the subsample for which data are available across all scenarios to factor out this confounder. However, that would have led to overall unstable results as our data shows. Another alternative would have been to use ICA to correct artifacts in the data. However, correcting artifacts rather than removing them with ICA can lead to spurious connectivity patterns, thereby limiting comparisons among the different features examined here. Nevertheless, since our work points to an almost general advantage of artifact removal for biometric purposes, future work should test different artifact removal methods to determine the most favorable approach and specifically compare artifact-free data with artifact-corrected data.

Template creation in the present approach relied on averaging across several trials. In contrast, EEG biometrics should work efficiently at the single-trial level [36]. However, this criticism was mainly targeted at situations where the ERP is used as a basis to extract EEG biometric features, which requires at least 20, ideally more than 100 repetitions. We averaged over 12, 8, or 3 repetitions for the three different conditions, which is relatively feasible. Our motivation for averaging was that it is intended to reduce the likelihood that templates rely on outliers in the data, which can easily arise from noise or artifacts. However, while we did not systematically investigate this, the effect seemed to be minor.

While we excluded artifacts, we corrected ocular movements rather than excluding contaminated data segments. Prior research often relied on eyes-closed data to avoid the confounding ocular

artifacts. Please note that this would not have been possible since the tasks were administered with visual instructions to participants. If we had excluded all artifacts, including ocular artifacts, the amount of data available would have been too low. A systematic variation in including vs. excluding different types of artifacts from the data (ocular, muscular, movement, and sweating), with provoked artifacts, is recommended for future research. Specifically, the natural occurrence of artifacts in this work prevented us from systematically investigating ocular artifacts, which are too frequent to be excluded. Therefore, we suggest that future work should systematically induce ocular and muscular artifacts to compare the differential effects on biometric performance and the stability of the extracted features when keeping these artifacts (ocular, muscular, or both) and when excluding them.

An additional limitation concerns the restricted diversity of cognitive tasks. We tested only three conditions (verbal learning/word pairs, FT, and VR recognition). A broader range of cognitive tasks, including motor imagery and resting-state paradigms, would allow to compare the results to existing studies, which often rely on PhysioNet. Conducting the analysis with datasets that include additional cognitive tasks would allow for systematic assessment of the generalizability of the presented task-dependent effects of artifact exclusion. Recent deep learning approaches to EEG biometrics increasingly explore this task diversity systematically within the context of task-invariant representations [72, 73]. However, datasets that provide both a large number of individuals and a large variety of tasks are still rare.

A further limitation concerns the artifact exclusion criteria themselves. We relied on fixed, predefined voltage-based thresholds, as implemented in Brain Vision Analyzer, to flag and exclude contaminated segments. While these criteria are well-established, highly practical, and widely used, fixed thresholds may fail to detect subtler artifacts that nonetheless distort biometric features. Adaptive or data-driven artifact detection, including recent deep-learning-based approaches [74], may offer a more nuanced alternative and should be compared systematically against threshold-based exclusion in future work.

Additionally, it should be noted that participant numbers declined substantially across the six sessions (Table 2), most markedly for the more demanding word-pair and VR tasks. This progressive dropout reduces statistical power in the later sessions and limits the comparability of estimates obtained early versus late in the protocol; results for sessions 5 and 6 in particular should be interpreted with caution given the smaller and potentially nonrepresentative subsamples retained at that point.

Finally, the sample includes mainly individuals with epilepsy, which were shown to exhibit greater distinctiveness and reliability in their EEG data [10, 75]. It is possible that the results would look different when the tests were carried out in a different population.

5 | Conclusion

Our results underscore the importance of considering whether to include or exclude artifacts, depending on the feature chosen for an EEG biometric system. Different cognitive tasks, alongside their different involvement of brain regions and, consequent to the experimental procedure, different lengths of individual data segments, lead to varying effects of artifacts in the data, but this effect also depends on the type of feature. We recommend the following:

1. In absence of further knowledge about the effect of artefacts on a specific feature, artefacts should be excluded. On average, artifact exclusion led to lower error rates of EEG biometrics in our dataset.
2. For frequency-dependent features, artifact exclusion should be considered, as it is likely to enhance biometric performance and reduce error rates. For frequency-independent features, artifacts in the data might carry additional information. We suggest that the role of artifacts should always be examined to confirm whether their additional value is present only if the artifacts are of human origin, for example, when artifacts are attributable to the individual as unique and stable patterns of muscular artifacts.
3. Further efforts should be directed toward designing a biometric system that is robust to artifacts, in favor of generating larger sample sizes for future studies and enhancing the system's usability.
4. EEG biometrics should no longer be evaluated within sessions, regardless of whether the data includes or excludes artifacts.
5. Small sample sizes are not a valid approach to assess EEG biometric performance. We need robust evidence generated in large samples. To this end, we need large, open-source EEG databases with multiple sessions from different days.

Author Contributions

Yvonne Höller: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing – original draft, writing – review and editing. **Arne C. Bathke:** supervision, writing – review and editing. **Andreas Uhl:** conceptualization, supervision, writing – review and editing.

Acknowledgments

We want to thank Eugen Trinka, Julia Höfler, Judith Dobesberger, Giorgi Kuchukhidze, Markus Leitinger, Margaritha Kirschner, Elisabeth Schmid, Daniela Sinadinoska, Susanne Grinzinger, Martin Scherr, and Wolfgang Staffen for recruitment of patients, and all participants for their time.

Funding

The data used for the presented research was acquired in a project funded by the Austrian Science Fund (FWF) (Grant T 798-B27) and by the Paracelsus Medical University Research Fund (PMU FFF) (Grant A-16/02/021-HÖL). The funders played no role in the present research, and no conflict of interest can be construed from the source of funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. H.-L. Chan, P.-C. Kuo, C.-Y. Cheng, and Y.-S. Chen, "Challenges and Future Perspectives on Electroencephalogram-Based Biometrics in Person Recognition," *Frontiers in Neuroinformatics* 12 (2018): 66.

2. F. Gondesen, M. Marx, and D. Gollmann, "EEG-Based Biometrics," in *Biometric-Based Physical and Cybersecurity Systems*, (Springer, 2019): 287–318.

3. R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "The Electroencephalogram as a Biometric," in *Canadian Conference on Electrical and Computer Engineering*, 2, (Canada: IEEE.

4. S. Marcel and J. D. R. Millan, "Person Authentication Using Brainwaves (EEG) and Maximum A Posteriori Model Adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 29, no. 4 (2007): 743–752.

5. M. Kostilek and J. Šťastný, "EEG Biometric Identification: Repeatability and Influence of Movement-Related EEG," in *2012 International Conference on Applied Electronics (AE 2012)*, (Pilsen, Czech Republic: IEEE, 2012), 147–150.

6. M. Wang, K. Kasmarik, A. Bezerianos, K. C. Tan, and H. Abbass, "On the Channel Density of EEG Signals for Reliable Biometric Recognition," *Pattern Recognition Letters* 147 (2021): 134–141.

7. E. Maiorana, D. La Rocca, and P. Campisi, "On the Permanence of EEG Signals for Biometric Recognition," *IEEE Transactions on Information Forensics and Security* 11, no. 1 (2016): 163–175.

8. T. W. Kjaer, L. S. Remvig, A. W. Helge, and J. Duun-Henriksen, "The Individual Ictal Fingerprint: Combining Movement Measures With Ultra Long-Term Subcutaneous EEG in People With Epilepsy," *Frontiers in Neurology* 12 (2021): 718329.

9. M. Del Pozo-Banos, J. A. B. Alonso, J. R. Ticay-Rivas, and C. M. Travieso, "Electroencephalogram Subject Identification: A Review," *Expert Systems With Applications* 41, no. 15 (2014): 6537–6554.

10. Y. Holler, A. C. Bathke, and A. Uhl, "Age, Sex, and Pathology Effects on Stability of Electroencephalographic Biometric Features Based on Measures of Interaction," *IEEE Transactions on Information Forensics and Security* 14, no. 2 (2019): 459–471.

11. G. Huang, Z. Hu, W. Chen, et al., "M3CV: A Multi-Subject, Multi-Session, and Multi-Task Database for EEG-Based Biometrics Challenge," *NeuroImage* 264 (2022): 119666.

12. O. Ozdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Adversarial Deep Learning in EEG Biometrics," *IEEE Signal Processing Letters* 26, no. 5 (2019): 710–714.

13. P. Cserti, B. Végső, G. Kozmann, Z. Nagy, F. De Vico Fallani, and F. Babiloni, "Methods to Highlight Consistency in Repeated EEG Recordings," *IFAC Proceedings Volumes* 45, no. 18 (2012): 23–27.

14. R. Palaniappan and K. V. R. Ravi, "A New Method to Identify Individuals Using Signals from the Brain," in *Proceedings of the 2003 Joint Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia*, (Singapore: IEEE.

15. G. K. Singhal and P. RamKumar, "Person Identification Using Evoked Potentials and Peak Matching," in *Biometrics Symposium*, (Baltimore, Maryland, USA: IEEE.

16. A. Yazdani, A. Roodaki, S. H. Rezatofighi, K. Misaghian, and S. K. Setarehdan, "Fisher Linear Discriminant Based Person Identification Using Visual Evoked Potentials," in *9th International Conference on Signal Processing*, (Beijing, China: IEEE.

17. R. Palaniappan, "Electroencephalogram Signals From Imagined Activities: A Novel Biometric Identifier for a Small Population," in *Intelligent Data Engineering and Automated Learning – IDEAL*, (Berlin, Heidelberg, Germany: Springer.

18. R. Palaniappan and D. P. Mandic, "EEG Based Biometric Framework for Automatic Identity Verification," *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* 49, no. 2 (2007): 243–250.

19. X. Bao, J. Wang, and J. Hu, "Method of Individual Identification Based on Electroencephalogram Analysis," in *International Conference on New Trends in Information and Service Science*, (Beijing, China: IEEE.

20. S.-K. Yeom, H.-I. Suk, and S.-W. Lee, "Person Authentication From Neural Activity of Face-Specific Visual Self-Representation," *Pattern Recognition* 46, no. 4 (2013): 1159–1169.

21. S. Sun, "Multitask Learning for EEG-Based Biometrics," in *19th International Conference on Pattern Recognition*, (Tampa, Florida, USA: IEEE.

22. M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "Permanence of the CEREBRE Brain Biometric Protocol," *Pattern Recognition Letters* 95 (2017): 37–43.

23. M. Näpflin, M. Wildi, and J. Sarnthein, "Test–retest Reliability of Resting EEG Spectra Validates a Statistical Signature of Persons," *Clinical Neurophysiology* 118, no. 11 (2007): 2519–2524.

24. D. La Rocca, P. Campisi, and G. Scarano, "On the repeatability of EEG features in a biometric recognition framework using a resting state protocol," in *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing (BIOSTEC 2013), Special Session MPBS*, 1, (Barcelona, Spain: SCITEPRESS.

25. B. C. Armstrong, M. V. Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, and S. Laszlo, "Brainprint: Assessing the Uniqueness, Collectability, and Permanence of a Novel Method for ERP Biometrics," *Neurocomputing* 166 (2015): 59–67.

26. R. Das, E. Maiorana, and P. Campisi, "EEG Biometrics Using Visual Stimuli: A Longitudinal Study," *IEEE Signal Processing Letters* 23, no. 3 (2016): 341–345.

27. W. Yunqi and L. Najafizadeh, "On the Invariance of EEG-Based Signatures of Individuality with Application in Biometric Identification," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Orlando, Florida, USA: IEEE.

28. K. Brigham and B. V. K. V. Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems*, (Washington, D.C., USA: IEEE.

29. S. Yang, F. Deravi, and S. Hoque, "Task Sensitivity in EEG Biometric Recognition," *Pattern Analysis and Applications* 21, no. 1 (2018): 105–117.

30. M. Fraschini, S. M. Pani, L. Didaci, and G. L. Marcialis, "Robustness of Functional Connectivity Metrics for EEG-Based Personal Identification Over Task-Induced Intra-Class and Inter-Class Variations," *Pattern Recognition Letters* 125 (2019): 49–54.

31. M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass, "Convolutional Neural Networks Using Dynamic Functional Connectivity for EEG-Based Person Identification in Diverse Human States," *IEEE Transactions on Information Forensics and Security* 14, no. 12 (2019): 3259–3272.

32. M. Wang, J. Hu, and H. A. Abbass, "BrainPrint: EEG Biometric Identification Based on Analyzing Brain Connectivity Graphs," *Pattern Recognition* 105 (2020): 107381.

33. M. R. F. Mota, P. H. L. Silva, E. J. S. Luz, et al., "A Deep Descriptor for Cross-Tasking EEG-Based Recognition," *PeerJ Computer Science* 7 (2021): e549.

34. Y. Di, X. An, W. Zhong, S. Liu, and D. Ming, "The Time-Robustness Analysis of Individual Identification Based on Resting-State EEG," *Frontiers in Human Neuroscience* 15 (2021): 672946.

35. A. Rahman, M. E. H. Chowdhury, A. Khandakar, et al., "Robust Biometric System Using Session Invariant Multimodal EEG and

Keystroke Dynamics by the Ensemble of Self-ONNs," *Computers in Biology and Medicine* 142 (2022): 105238.

36. R. Ashenaei, A. Asghar Beheshti, and T. Yousefi Rezaii, "Stable EEG-Based Biometric System Using Functional Connectivity Based on Time-Frequency Features With Optimal Channels," *Biomedical Signal Processing and Control* 77 (2022): 103790.

37. R. Plucińska, K. Jędrzejewski, U. Malinowska, and J. Rogala, "Leveraging Multiple Distinct EEG Training Sessions for Improvement of Spectral-Based Biometric Verification Results," *Sensors* 23, no. 4 (2023): 2057.

38. J. Xu, E. Zhou, Z. Qin, T. Bi, and Z. Qin, "Electroencephalogram-Based Subject Matching Learning (ESML): A Deep Learning Framework on Electroencephalogram-Based Biometrics and Task Identification," *Behavioral Sciences* 13, no. 9 (2023): 765.

39. E. A. Abdel-Ghaffar and M. Salama, "The Effect of Stress on a Personal Identification System Based on Electroencephalographic Signals," *Sensors* 24, no. 13 (2024): 4167.

40. H. Liu, X. Jin, D. Liu, W. Kong, J. Tang, and Y. Peng, "Joint Disentangled Representation and Domain Adversarial Training for EEG-Based Cross-Session Biometric Recognition in Single-Task Protocols," *Cognitive Neurodynamics* 19, no. 1 (2025): 31.

41. M. Fallahi, P. Arias Cabarcos, and T. Strufe, "Advancing Brainwave-Based Biometrics: A Large-Scale, Multi-Session Evaluation," *IEEE Transactions on Biometrics, Behavior, and Identity Science* 8, no. 3 (2026): 445–460.

42. M. Usman, N. Sultan, A. Nasim, B. Ayaz, J. Fatima, and F. Nosheen, "Balancing Noise Reduction and Neural Signature Preservation in EEG Biometrics," *Scientific Reports* 16, no. 1 (2026): 6674.

43. S. N. A. Seha and D. Hatzinakos, "A New Training Approach for Deep Learning in EEG Biometrics Using Triplet Loss and EMG-Driven Additive Data Augmentation," *Neurocomputing* 488 (2022): 194–211.

44. A. K. Chaurasia, M. Fallahi, T. Strufe, P. Terhörst, and P. A. Cabarcos, "Cabarcos P.a.: NeuroIDBench: An Open-Source Benchmark Framework for the Standardization of Methodology in Brainwave-Based Authentication Research," *Journal of Information Security and Applications* 85 (2024): 103832.

45. Y. Höller, S. G. Eyjólfsdóttir, M. Rusiňák, L. S. Guðmundsson, and E. Trinka, "Movement Termination of Slow-Wave Sleep—A Potential Biomarker?" *Brain Sciences* 14, no. 5 (2024): 493.

46. Y. Höller, S. Eyjólfsdóttir, F. J. Van Schalkwijk, and E. Trinka, "The Effects of Slow Wave Sleep Characteristics on Semantic, Episodic, and Procedural Memory in People With Epilepsy," *Frontiers in Pharmacology* 15 (2024): 1374760.

47. S. G. Eyjólfsdóttir, E. Trinka, and Y. Höller, "Shorter Duration of Slow Wave Sleep Is Related to Symptoms of Depression in Patients With Epilepsy," *Epilepsy and Behavior* 149 (2023): 109515.

48. F. J. van Schalkwijk, W. R. Gruber, L. A. Miller, E. Trinka, and Y. Höller, "Investigating the Effects of Seizures on Procedural Memory Performance in Patients With Epilepsy," *Brain Sciences* 11, no. 2 (2021): 261.

49. Y. Höller, C. Höhn, F. Schwimmbeck, G. Plancher, and E. Trinka, "Effects of Antiepileptic Drug Tapering on Episodic Memory as Measured by Virtual Reality Tests," *Frontiers in Neurology* 11 (2020): 93.

50. G. Gratton, M. G. H. Coles, and E. Donchin, "A New Method for Off-Line Removal of Ocular Artifact," *Electroencephalography and Clinical Neurophysiology* 55, no. 4 (1983): 468–484.

51. S. L. Marple, *Digital Spectral Analysis With Applications* (Prentice-Hall, 1987).

52. R. E. Greenblatt, M. E. Pflieger, and A. E. Ossadtchi, "Connectivity Measures Applied to Human Brain Electrophysiological Data," *Journal of Neuroscience Methods* 207, no. 1 (2012): 1–16.

53. A. Schlogl and C. Brunner, "BioSig: A Free and Open Source Software Library for BCI Research," *Computer* 41, no. 10 (2008): 44–50.

54. A. Schlögl, "A Comparison of Multivariate Autoregressive Estimators," *Signal Processing* 86, no. 9 (2006): 2426–2429.

55. V. K. Murthy, "Estimation of the Cross-Spectrum," *The Annals of Mathematical Statistics* 34, no. 3 (1963): 1012–1021.

56. M. Eichler, "On the Evaluation of Information Flow in Multivariate Systems by the Directed Transfer Function," *Biological Cybernetics* 94, no. 6 (2006): 469–482.

57. G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying True Brain Interaction From EEG Data Using the Imaginary Part of Coherency," *Clinical Neurophysiology* 115, no. 10 (2004): 2292–2307.

58. W. Gersch and G. V. Goddard, "Epileptic Focus Location: Spectral Analysis Method," *Science* 169, no. 3946 (1970): 701-702.

59. L. A. Baccala, K. Sameshima, and D. Y. Takahashi, "Generalized Partial Directed Coherence," in *15th International Conference on Digital Signal Processing*, (Cardiff, Wales, UK: IEEE.

60. M. J. Kaminski and K. J. Blinowska, "A New Method of the Description of the Information Flow in the Brain Structures," *Biological Cybernetics* 65, no. 3 (1991): 203–210.

61. A. Korzeniewska, M. Mańczak, M. Kamiński, K. J. Blinowska, and S. Kasicki, "Determination of Information Flow Direction Among Brain Structures by a Modified Directed Transfer Function (dDTF) Method," *Journal of Neuroscience Methods* 125, no. 1-2 (2003): 195–207.

62. J. F. Geweke, "Measures of Conditional Linear Dependence and Feedback Between Time Series," *Journal of the American Statistical Association* 79, no. 388 (1984): 907–915.

63. S. L. Bressler, C. G. Richter, Y. Chen, and M. Ding, "Cortical Functional Network Organization From Autoregressive Modeling of Local Field Potential Oscillations," *Statistics in Medicine* 26, no. 21 (2007): 3875–3885.

64. M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating Causal Relations in Neural Systems: Granger Causality, Directed Transfer Function and Statistical Assessment of Significance," *Biological Cybernetics* 85, no. 2 (2001): 145–157.

65. E. Maiorana, D. L. Rocca, and P. Campisi, "Eigenbrains and Eigentensorbrains: Parsimonious Bases for EEG Biometrics," *Neurocomputing* 171 (2016): 638–648.

66. K. V. R. Ravi and R. Palaniappan, "Leave-One-Out Authentication of Persons Using 40 Hz EEG Oscillations," in *The International Conference on Computer as a Tool*, (Belgrade, Serbia and Montenegro: IEEE.

67. R. Palaniappan and K. V. R. Ravi, "Improving Visual Evoked Potential Feature Classification for Person Recognition Using PCA and Normalization," *Pattern Recognition Letters* 27, no. 7 (2006): 726–733.

68. P. Tangkraingkij, C. Lursinsap, S. Sanguansintukul, and T. Desudchit, "Selecting Relevant EEG Signal Locations for Personal Identification Problem Using ICA and Neural Network," in *Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science: June*, (Shanghai, China: IEEE/ACIS, 2009), 616–621.

69. Y. Höller, C. Helmstaedter, and K. Lehnertz, "Quantitative Pharmaco-Electroencephalography in Antiepileptic Drug Research," *CNS Drugs* 32, no. 9 (2018): 839–848.

70. Y. Höller and A. Uhl, "Do EEG-Biometric Templates Threaten User Privacy?" in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, (Innsbruck, Austria: ACM.

71. R. Damaševičius, R. Maskeliūnas, E. Kazanavičius, and M. Woźniak, "Combining Cryptography With EEG Biometrics," *Computational Intelligence and Neuroscience* 2018 (2018): 1867548.

72. A. E. Albaiati, M. F. Akbar, M. D. Hssayeni, et al., "Deep Learning Approaches for EEG-Based Biometrics: A Systematic Review," *IEEE Access* 13 (2025): 171025–171047.

73. Y. Gong, M. Wang, Y. Zhang, et al., "A Unified Deep Learning-Based EEG Biometric Authentication System for Cross-Session Scenarios," in *Advanced Data Mining and Applications*, 1st ed., (Springer, 2024): 48–62.

74. M. Wang, X. Yin, and J. Hu, "Cancellable Deep Learning Framework for EEG Biometrics," *IEEE Transactions on Information Forensics and Security (TIFS)* 19 (2024): 3745–3757.

75. Y. Höller, K. Butz, and A. Thomschewski, "Reliability of EEG Interactions Differs Between Measures and Is Specific for Neurological Diseases," *Frontiers in Human Neuroscience* 11 (2017): 350.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Supporting Information.** Figures S1–S15 show the subselection results for individual features. Selection likelihood across cross-validation scenarios is indicated as a percentage of how often a feature was selected, which is illustrated in heat-map color coding (see the colormap legend, normalized to the interval 0 to 1). Figure S1: Af. Figure S2: AR. Figure S3: COH. Figure S4: DC. Figure S5: dDTF. Figure S6: DTF. Figure S7: ffDTF. Figure S8: GGC. Figure S9: GPDC. Figure S10: hh. Figure S11: pCOH. Figure S12: PDC. Figure S13: PDCF. Figure S14: PSD. Figure S15: S.