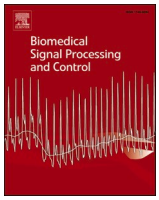
Biomedical Signal Processing and Control 76 (2022) 103664

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Mean curve length: An efficient feature for brainwave biometrics

Reza Yahyaei, Tolga Esat Özkurt *

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Keywords:
Biometrics
Connectivity
EEG
Neural oscillations
Nonlinearity
Resting-state

ABSTRACT

Electroencephalography (EEG) as a biometric modality has gained considerable interest in recent years. Many state-of-the-art methods have focused on increasing the recognition accuracy. However, the more complex and manipulative the methods become, the less practical and generalized they are in real-life applications. In this study, we prioritized computational efficiency and evaluated the model performance. In this direction, we propose the mean curve length (MCL), a simple measure quantifying signal complexity, which is analytically and empirically related to the Katz fractal dimension. By merely being the average of the absolute value of the first-order difference of a signal, MCL is arguably the most computationally efficient feature that can be extracted from an EEG signal. In this paper, we utilized it for person identification and authentication on a large standard dataset comprising 109 subjects under the eyes-open (EO) and eyes-closed (EC) resting state conditions. We employed a Mahalanobis distance-based classifier both for identification and authentication tasks. Our results indicate that in addition to its simplicity and low computational cost, MCL provides a remarkably high individual distinction as well. Specifically, recognition accuracies were 99.4% (EO) and 98.8% (EC) for identification, and for authentication, equal error percentages of 6.33% (EO) and 10.50% (EC) were obtained. Our study offers a fast and accurate neural biometric recognition scheme promising especially for practical real-world and real-time applications. It further proves the effectiveness of nonlinear signal measures in individual discrimination, and promotes shifting the focus beyond the conventional brain oscillatory and connectivity measures commonly fostered in EEG-based biometrics literature.

1. Introduction

Human brain contains the most intimate thoughts, memories and feelings of individuals, and hence carry firsthand information of personal identities. This bestows a unique potential capability upon biometrics based on scalp electroencephalography (EEG). Being a continuous data source, EEG requires the subject to be present, alive and in a proper mental state to be valid, making it robust against identity theft and forgery. Based on the seven criteria defined in [1] — universality, uniqueness, permanence, collectability, performance, acceptability and circumvention — brainwaves hold medium to high scores in comparison to the other biometrics such as those obtained from DNA, face, fingerprint and iris [2]. Cutting-edge mobile, wearable and wireless EEG systems with dry electrodes allow noninvasive real-time acquisition of brain signals by setting free any constraint on the movements of the subjects in their environment [3]. This surge in technology grants practicality as well as robustness to EEG systems for biometric recognition from the brain.

There have been a wide range of studies on EEG biometrics in terms of features, classifiers, frequency bands, channel combinations and signal acquisition protocols [2]. The features that have been used in the state-of-the-art EEG-based biometrics can be divided into three categories: univariate measures such as power spectral density (PSD) [4–6] and autoregressive (AR) model coefficients [7–9], bivariate functional connectivity measures [10] such as spectral coherence [6] and mutual information [11], and multivariate network measures including weighted nodal degree [12] and eigenvector centrality [13]. Amongst all, features based on brain oscillatory characteristics (spectral band activities), specifically the spectral connectivity measures that hold a superior performance [6,14], have been preferred in most of the studies, while features characterizing signal complexity and nonlinearity were rarely used. The studies employing signal complexity based features have used entropy measures [15–17], fractal dimension [17–19] and $1/f$ component of background PSD [20], being related to the fractal dimension [21].

Here, we propose a plain and computationally efficient measure, the

* Corresponding author.
E-mail addresses: ozkurt@metu.edu.tr, tolgaozkurt@gmail.com (T. Esat Özkurt).

https://doi.org/10.1016/j.bspc.2022.103664
Received 3 January 2022; Received in revised form 23 February 2022; Accepted 19 March 2022

mean curve length (MCL), for EEG-based biometrics intimately related to the Katz fractal dimension (KFD) [22]. It is defined as the average of the absolute value of the first-order difference of a signal, making it one of the most computationally efficient features ever possible to be extracted from EEG signals. MCL is also the mean of curve length (CL), which was proposed and adopted as a feature for detection and prediction of seizures in epileptic patients [22–24]. Özkurt et al. [25] used a generalized form of it, called the short-time average magnitude difference function (AMDF), for seizure detection, and compared its performance with conventional CL. For the case of AMDF, the difference can be computed with a selected delay instead of unity as in CL, which enables detection of changes in an extended number of frequencies. Because of this capability, AMDF has been used in speech signal processing for time-delay estimation and pitch detection [26–27]. A short-time AMDF was introduced by [28] in order to capture sleep spindle related regularities from sleep EEG signals. In [29], a more generalized form of AMDF was defined by adding an integer-wise degree in order to identify nonlinear properties with an application to Parkinson's disease subcortical and cortical signals. Moreover in the same study, those generalized AMDF values were shown to be correlated with clinical motor symptoms.

Classification methods for EEG-based biometrics have been realized under two major paradigms: i) conventional classifiers, in particular those based on distance (e.g., Euclidean) or similarity (e.g., Pearson's correlation coefficient) [30–33] to some pre-stored user templates and ii) deep learning solutions, where convolutional neural networks (CNN) have often been employed [34–37]. While deep learning based classifiers automatically extract deep representations of features and generally yield a superior performance, the gain in performance comes with a computational cost in turn. Moreover, instead of explicit templates, user data are implicitly stored within the trained network parameters. Hence, adding or removing a user requires the network model to be retrained. However in a template-based scheme, since the user records are explicit, one can easily add or remove them from the database. As the classification only involves finding distances or similarity measures in reference to the templates, little to no training is required for the classifier. Due to the aforementioned reasons, a Mahalanobis distance-based classifier was chosen to be employed in our study. Mahalanobis distance provides a computationally low-cost classifier that produces closed-form solutions, without any requirement for parameter tuning or optimization. There are limitations in computing the Mahalanobis distance for high-dimensional features such as functional connectivity feature matrices. To address this problem and improve the overall classification performance, a modified version of the Mahalanobis distance was implemented: the implicit data transformation within the definition of Mahalanobis distance was carried out explicitly, a dimensionality reduction step was added to the transformation and subsequently the distance was computed.

In this study, MCL was operated as a feature for EEG-based biometric identification and authentication on a large standard dataset. For the purpose of comparison, we also included KFD and some other highly discriminative features reported in the literature. Mahalanobis distance and a modified implementation for high-dimensional features were employed for classification.

2. Materials and methods

2.1. Dataset and preprocessing

This study benefited from the publicly available PhysioNet EEG Motor Movement/Imagery Dataset [38], a frequently utilized database in EEG biometric studies [4–6,9,10,13,15,20,30,32,34–36,39–44]. It comprises EEG recordings from 109 participants in 14 experimental runs, including two baseline runs of resting state conditions with eyes-open (EO) and eyes-closed (EC), both lasting one minute. As EEG is known to be sensitive to human mental states [42,44], spontaneous brain data acquired during resting state have been mostly preferred for biometric recognition to keep the intra-individual variations in brain activity as low as possible. Thus in the current study, the EO and EC resting state data were used for the evaluation of the proposed methodology.

Data were recorded using a BCI2000 system (http://www.bci2000.org) with 64 electrodes and a sampling rate of 160 Hz. Electrodes were arranged according to the international 10–10 system. All EEG signals were referenced to the average of the signals coming from earlobes.

For each subject and condition, the recorded signals were split into six non-overlapping segments of 10 s to generate samples, five of them for training the model and one for testing it. A lowpass FIR filter with a cut-off frequency of 50 Hz was applied to each segment separately to eliminate the noise-ridden higher frequency components, including the 60 Hz power line frequency and to retain the targeted dominant brain rhythmic activity.

2.2. Feature extraction

The main focus of this EEG-based biometric study is the mean curve length (MCL) that can be computed only by a summation operation applied on its discrete derivative (defined in Section 2.2.1). Its analytical simplicity enables great computationally efficiency, which is a desirable quality for implementing a practical biometric system. We also included several other features mainly for the purpose of comparison with MCL on the same dataset, particularly in terms of biometric recognition performance and computational efficiency. They comprise the highly discriminative features that have been commonly preferred in EEG biometric studies thus far, along with two popular fractal dimension estimators.

In Section 2.2.2, we demonstrate what common characteristics MCL shares with a fractal dimension estimator, called the Katz Fractal Dimension (KFD). We also show their empirical similarity from their correlation using EEG data. In Section 2.2.3, we added up two other signal-complexity-based features: Higuchi's fractal measure (HM) [45] and the aperiodic component of the power spectrum, i.e., the power-law exponent and offset of the $1/f$ background trend. The former is a popular fractal dimension estimator that has been used quite frequently in recent EEG signal pattern recognition studies for various cognitive and clinical applications (e.g., [46–48]). The latter feature, $1/f$ power-law exponent γ is linearly related to the fractal dimension $\left(D = \frac{5-\gamma}{2}\right)$, which has been recently applied in EEG-based biometrics by [20]. The remaining features for comparison were chosen amongst the most commonly utilized EEG metrics of spectral band power and connectivity. These two feature types are supposed to quantify functional segregation and functional

Table 1
Computational complexities of MCL and other features in terms of feature vector size, average number of principal components over subjects and conditions (only for PLV, COH and AR), average runtime over subjects and conditions and analytical time complexity (e = # of channels, t = # of time points, k = k_{max} maximum scale parameter for HM, p = order of the AR model). Features marked by † were incorporated with a dimensionality reduction corresponding to a 99% retained variance in their classification.

Features	Feature size (→Average # of principal components)	Average runtime	Complexity
MCL	64	0.03 s	O(e × t)
KFD	64	0.07 s	O(e × t)
HM	64	0.38 s	O(e × k^2 × t)
PSD	64 × 2	0.29 s	O(e × t)
AP	64 × 2	115.6 s	O(e × t^2)
PLV†	(64 × 63 / 2) × 2 → 250p.c.	2.3 s	O(e^2 × t)
COH†	(64 × 63 / 2) × 2 → 163p.c.	1.17 s	O(e^2 × t)
AR†	64 × 5 → 56p.c.	0.14 s	O(e × p × t)

integration of the brain [49], respectively.

Table 1 exhibits the superior computational efficiency of MCL with respect to the aforementioned features. Specifically, it is more than twice faster than KFD. It is expectedly much faster than HM (more than 12 times faster in our case) as the latter requires the computation of curve length for multiple times (quadratically proportional to the parameter of k_{max}, see Table 1) in different scales and fitting a straight line over them, which make it inherently more complex than MCL by definition. MCL is also two orders of magnitude faster than phase-locking value and spectral coherence, as it has a linear computational complexity with respect to the number of channels, opposed to the quadratic complexity of the functional connectivity metrics.

2.2.1. Mean curve length (MCL)

Curve length (CL) was originally proposed by Esteller et al. [22] as a computationally efficient substitute for the Katz fractal dimension [50]. As the name suggests, curve length is the sum of the linear distances between successive points on a curve. For the case of EEG, the so-called curve is basically equivalent to the signal trace. Since the time-axis contributes equally to the distances for a fixed sampling rate, it is not taken into account while computing CL. Hence, CL is the sum of the absolute value of the first-order finite difference (discrete derivative) for a time-series. Mean curve length (MCL) is simply obtained by averaging CL:

$$MCL = \frac{1}{N-1}\sum_{n=0}^{N-2}|\Delta_1 x(n)| = \frac{1}{N-1}\sum_{n=0}^{N-2}|x(n+1)-x(n)| \quad (1)$$

where N is the signal length and $n = 0, 1, \ldots, N\text{-}1$. Fig. 1 exhibits a visual illustration of MCL.

2.2.2. Derivation of MCL from Katz fractal dimension (KFD)

Katz [50] introduced a formulation for the fractal characterization of waveforms, which are two-dimensional curves that monotonically move forwards in one dimension (e.g., the time dimension). This can be especially useful for analyzing and comparing complex waveforms such as EEG signals. The fractal dimension D as defined by Katz [50] is given by.

$$D = \log(L)/\log(d) \quad (2)$$

where L is the curve length in two dimensions and is defined as.

$$L = \sum_{n=0}^{N-2}\sqrt{(x(n+1)-x(n))^2 + \Delta t^2}, \quad (3)$$

where Δt is the constant unit of time (i.e., the sampling period) and d is the curve diameter, defined as the farthest distance between the starting point and any other point on the curve:

$$d = \max\sqrt{(x(n)-x(1))^2 + (t(n)-t(1))^2} \quad (4)$$

for $n = 0, 1, \ldots, N\text{-}1$. Notice that unlike MCL, the time-axis has to be taken into account for computing the Katz fractal dimension (KFD), since it does not equally contribute to the distances in Eq. (4).

The ratio in Eq. (2) compares the actual length of the curve with the minimum length required to create a curve of the same spatial extent. Complex and convoluted curves tend to have higher fractal dimensions, since they can fit a long length within a small extension of space, i.e., they fill more of the space.

Since arbitrary units of length for the definition in Eq. (2) produce different results, a standard unit denoted as a, describing the average distance between the successive points, was defined in order to normalize KFD [50]:

$$D = \frac{\log(L/a)}{\log(d/a)} = \frac{\log(m)}{\log(m)+\log(d/L)} \quad (5)$$

where $m = L/a$ is the number of steps on the curve. Despite this normalization, the difference between the units of x- and y-*axis* is not taken into account for a typical time-series. This is important, because by considering a time-series as a geometrical object, different values of fractal dimension can still be generated by choosing arbitrary scales for the time and amplitude axes.

The unit of time Δt has been treated differently in various implementations of KFD. For instance, in one implementation [51], Δt was taken as unity, while in another [52], the time dimension was altogether ignored. It is helpful to understand the effect of this parameter by considering the extreme cases. When the unit of time goes towards the infinity, the relative size of signal amplitude becomes so negligible that any curve stretches into a virtually straight line with a fractal dimension of approximately one. In the other extreme when the unit of time is zero (i.e., KFD is computed in one dimension), the curve diameter would merely be the largest peak in a signal. This would lead to inconsistent results because of the inevitable variations in signal amplitude, particularly caused by noise and the elimination of the normalization for signal length. In our own implementation, Δt was set to one to have a close but smaller magnitude than the average distance along the amplitude-axis (Eq. (3)) so that the effect of the latter remains dominant.

Katz's formulation of fractal dimension has an ambiguity with regard to time-series such as EEG signals, since the distances along the two dimensions have different units. In addition to that, some combinations of values for the parameters in Katz's formulation (i.e., m, L and d in Eq. (5)) lead to inconsistent results [22]. To overcome these problems, Esteller et al. [22] proposed that the logarithms in Katz's formulation could be dropped to obtain a more computationally efficient feature without compromising the detection capability. After eliminating the logarithmic functions in the numerator and denominator, Eq. (2) becomes L/d. Here, d can be considered as a normalization factor, which does not change significantly among time segments of equal length, and hence can be taken as a constant. With d being a constant, L is no longer needed to be calculated in two dimensions (see Section 2.2.2), thus it would be equivalent to CL. Finally, substituting the number of steps n for the curve diameter d leads to the definition of MCL in Eq. (1). MCL is a well-defined measure computed in one dimension, avoiding the problems of possibly singular or improper outputs [22] and dimensional unit inconsistency existent in KFD. We plotted MCL vs. KFD (Fig. 2) using the samples from all subjects in order to demonstrate the strong correlation (Pearson's $r = 0.96$, $p < 10^{-15}$) between these measures empirically.

2.2.3. Other features

The other features employed in this study are as follows: Higuchi's measure (HM) [45], power spectral density (PSD) [5], aperiodic

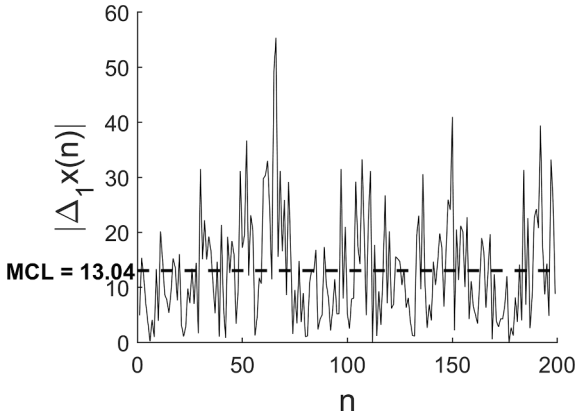


Fig. 1. An exemplary illustration of mean curve length that can be considered as the mean of absolute value of first-order finite difference (discrete derivative) of a signal.

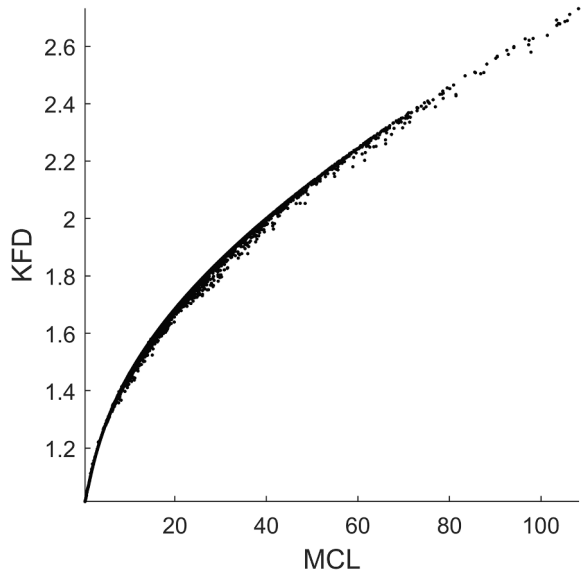


Fig. 2. KFD and MCL has high empirical correlation ($r = 0.96$, $p < 10^{-15}$). Samples (109 × 6 × 64) were pooled from all subjects, segments and channels.

component (AP) of PSD [20] and autoregressive (AR) coefficients [9]. In addition, we included two common cross-channel metrics of functional connectivity that quantify phase synchronization: phase-locking value (PLV) [42] and spectral coherence (COH) [6]. It has been shown that pairwise connectivity measures between channels may yield a better biometric recognition than single-channel spectral measures [6].

HM is known to be one of the popular fractal dimension estimators used in EEG classification studies [53–55]. In a comparative study of the fractal dimension estimators by Esteller et al. [56], they show that although HM yields the most accurate fractal dimension estimates for synthetic data, it is worse and less consistent than KFD in discriminating the brain states of epileptic patients. In another study for patients with encephalopathy [47], HM was reported to have a better discrimination ability than KFD. Hence, these studies suggest how the use and performance of fractal dimension measures are very much dependent on the specific data and application underhand.

It is worth noting that HM does not correspond to the curve length as such, rather, it corresponds to the change of rate (slope) of the curve lengths in different scales for $k = 1,\ldots,k_{\max}$. That's why it empirically correlated neither with MCL ($r = 0.19$), nor with KFD, strongly ($r = 0.15$; see Fig. S1 given in the Supplementary material). The discrepancy between KFD and HM can also be sensed from the results reported by [56]. The choice of $k_{\max}$ is said to matter for the fractal dimension estimation [46]. Thus, we implemented and tried HM for $k_{\max}$ varying between 4 and 16 and chose $k_{\max} = 6$, which provided a relatively higher performance. Although, for our case, the performance did not change much with respect to $k_{\max}$.

In the EEG biometric studies that took advantage of the frequency band activities, higher frequency bands, beta and gamma in particular, were shown to contain the most distinctive information about individuals [10,15,30], and were found to be the most robust bands for short-time data segments in a recent study [14]. Therefore, we extracted the PSD, PLV and COH features within the ranges of beta (12–30 Hz) and gamma (30–45 Hz) bands from the analytic signals, which were obtained by applying the Hilbert transform to the bandpass filtered signals. For each feature set, the estimates from these two bands were concatenated to construct a single feature vector.

PSD features were computed by taking the total power of the analytic bandpass filtered signals for each channel, and functional connectivity measures of PLV and COH were estimated between all pairs of channels (64 × 63/2 = 2016) [6,42]. Phase-locking value measures the consistency of the instantaneous phase difference between two signals, while spectral coherence also factors in the amplitudes of the signals in addition to their phase difference. Both PLV and COH range from zero to one, gradually indicating from non-synchrony to perfect synchrony between two signals.

AP features comprise the parameters of the power-law exponent linearly related to the fractal dimension and the offset. Both were estimated from a *1/f* model fitted on the logarithmic scale of the power spectrum. The model for the aperiodic component is defined as $L(f) = b - \log(\kappa + f^{\gamma})$, where *f, b,* γ and κ denote frequency, offset, the power-law exponent and the knee parameter, respectively. We estimated b and γ using the *Fitting Oscillations and One Over F* (FOOOF) toolbox [57], which was specifically developed for the analysis of periodic and aperiodic properties of electrophysiological signals. First, the power spectrum of each data segment was estimated by Welch's averaged periodogram method. As each segment was 10 s long, a Hanning window of 2 s with 50 % overlap was applied. Subsequently, the estimated power spectral values were passed to FOOOF to obtain the parameters of power-law exponent and offset. The frequency range for curve fitting was set to 1–40 Hz.

AR features were estimated using Burg's method for its lower computational complexity and better performance in comparison to the other AR parameter estimation techniques [2,8]. A model order of five was selected based on an EEG biometric study [9] which used the same dataset. Please note that all feature extraction procedures were performed using MATLAB (R2019b, The Mathworks, Inc., Natick, MA) on an Intel(R) Xeon(R) CPU E5- 2650 0 @ 2.00 GHz–2.00 GHz processor.

2.3. Classification

Mahalanobis distance-based classifier was employed both for identification and authentication tasks, as it yields high performance [58], low computational cost and scores (distances) enabling applicability for authentication. Computing Mahalanobis distance requires the estimation of the inverse covariance matrix for each class (subject) distribution. Since the number of samples for each class is small (five samples), the individual covariance matrices cannot be robustly estimated. To address this, we followed a common procedure consisting in approximating the covariance matrix of each class to be equal to the covariance matrix computed from all classes [6], thus leading to only one covariance matrix stored in the database. Following this approach, for the n features of each feature set, an ($n \times n$) covariance matrix was computed using 5 × 109 samples (5 segments and 109 subjects). After acquiring the inverse covariance matrix $\sum^{-1}$, the square Mahalanobis distance $d_{u,n}^2$ between an unidentified observation's feature vector η_u and a class centroid μ_n was calculated as.

$$d_{u,n}^2 = (\eta_u - \mu_n)^T \Sigma^{-1} (\eta_u - \mu_n) \tag{6}$$

where $u, n \in \{1, 2, \ldots, C\}$, with C denoting the number of classes. Class centroids are the subject templates stored in the database during user registration. They are the feature vectors that are obtained by averaging the five sample feature vectors (corresponding to the five segments designated for training, as described in Section 2.1) for each subject.

The observation is then predicted to belong to the class distribution with the minimum distance:

$$\hat{n} = argmin_n d_{u,n}^2 \tag{7}$$

For authentication, a distance threshold was admitted, such that the observations within that distance were accepted as belonging to the authorized person, while the ones farther than that were rejected as impostors. An appropriate threshold can be determined by taking into account the trade-off between false acceptance/rejection rates (see Section 2.4).

Mahalanobis classifier suffers from a deficiency with regard to high-dimensional features such as PLV and COH, as the covariance matrix can

become extremely large, possibly leading to a memory crash in a basic processor. Therefore, we modified the implementation of Mahalanobis distance in order to incorporate a dimensionality reduction. This was achieved by taking advantage of the fact that Mahalanobis distance is equivalent to Euclidean distance for a whitened distribution [59], thus allowing it to be computed in two separate steps: whitening the data and calculating the Euclidean distance. Whitening consists of decorrelating and standardizing features. Less important features, i.e., those with the least variances can be eliminated after decorrelating the data and before standardizing them (Fig. 3). We performed these steps via principal component analysis (PCA) and subsequently computed the Euclidean distances between the transformed feature vectors. This not only improved the computational efficiency by reducing the size of the feature vectors, but also increased the classification performance by helping to avoid the curse of dimensionality.

For the features of AR, PLV and COH, a dimensionality reduction step was added to the whitening process, such that the retained components (i.e., the transformed features) contained 99 % of the variance of the original features. PCA can be a much faster alternative to deep learning [42] or exhaustive feature selection [6] while dealing with high-dimensional functional connectivity features.

All classification procedures were implemented using the Scikit-learn library (https://www.scikit-learn.org) in Python 3. Codes for feature extraction and classification are available at https://github.com/RezaYahyaei/Paper2022.

2.4. Evaluation

For the evaluation of the proposed biometric system, we employed the standard metrics used in EEG-based biometrics literature [2]. Correct recognition rate (CRR), which represents the fraction of correct identifications, was used to assess identification performances. CRR values were computed in a leave-one-out cross-validation fashion as follows: from the six segments corresponding to each subject, one was leaved out for testing the model and the remaining five segments were used for its training. This was carried out for all six combinations of train/test set split. Subsequently, the mean and standard deviation of the CRR values were computed over the six cross-validation iterations.

Authentication performances are expressed in terms of the detection error trade-off (DET) curve, which depicts the trade-off between false rejection rate (FRR) and false acceptance rate (FAR) as a function of threshold. FRR is the fraction of rejections over all the access attempts by genuine users and FAR is the fraction of acceptances over all the access attempts by impostors. The point on a DET curve where the FRR and FAR values intersect is defined as the equal error rate (EER), which is a numeric metric of performance for authentication systems. Lower EER indicates better performance.

To construct a DET curve, all pairwise distances are needed, together with the labels indicating whether they are intra-class (genuine match) or inter-class (impostor match) distances. This was realized within each cross-validation iteration, where all pairwise distances between the test set and the templates (generated during the training) were stored. The test set and the template set both consisted of 109 elements corresponding to the subjects. Thus for 6 runs, there were 6×109 genuine and $6\times(109 \times 108)$ impostor matches in total.

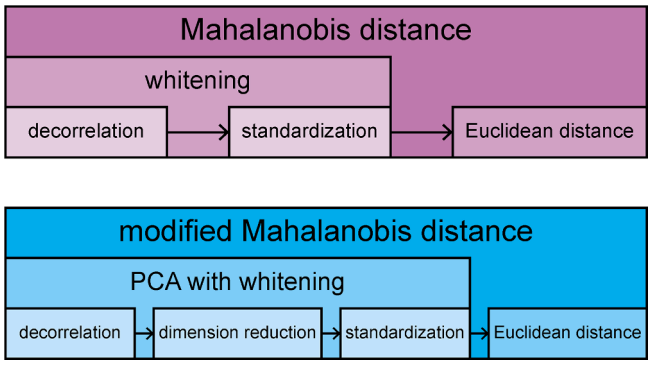


Fig. 3. Mahalanobis distance and its modified implementation.

3. Results

3.1. Identification and authentication performance

We evaluated the identification and authentication performances for MCL and the other included features. Identification was realized by assigning the observed data to the nearest existing class, while authentication was based upon deciding whether an observation belongs to a class or not, with respect to a distance threshold. A dimensionality reduction was incorporated into the classification pipeline for the features of PLV, COH and AR. The obtained performances per each feature are given in Table 2 in terms of the mean CRR $\pm$ the standard deviation of a 6-fold cross-validation for identification and EER for authentication.

All features yielded comparably high performances of over 95 %. Specifically, a CRR of 99.4 $\pm$ 0.4 % in the EO condition and a slightly lower CRR of 98.8 $\pm$ 1.1 % in the EC condition were obtained for MCL, which were close to the CRRs obtained for KFD. Even though KFD and HM did not correlate well as fractal dimension estimators (see Section 2.2), their identification and authentication performances were on a par. In general, the EO condition facilitated better identification performances than the EC condition, which is in line with literature [13,15,20,42].

Authentication performance was distinctively low only for the features of PSD and AP, with high EERs of more than 15 % (Fig. 4). For MCL, acceptable EERs of 6.29 % and 10.40 % were obtained, respectively for the EO and EC conditions.

The conspicuous difference of within-subject and between-subject variances for all channels exhibited in Fig. 5 bears witness to the high biometric performances achieved with MCL. This is due to the unique correspondence of subjective brainwaves to MCL traces being almost indistinguishable for temporally different segments of the EEG signals as shown in Fig. 6.

Please note that we tried out two conventional classifiers of multilayer perceptron (MLP) neural network and support vector machine (SVM) to see how they compare with the suggested Mahalanobis distance-based classifier identification results. MLP yielded very similar and comparable performance, however it takes more computational time for training and testing than the Mahalanobis distance-based classifier does. SVM had a similar run time with Mahalanobis distance-based classifier, but its identification accuracies were lower. The details of implementation, classifier parameters and identification performances can be found out in the Supplementary Material.

Table 2
Identification and authentication performances of MCL and five other features in terms of CRR (mean $\pm$ standard deviation of 6-fold cross-validation) and ERR (corresponding to the DET curves in Fig. 4), respectively. Features marked by † were incorporated with a dimensionality reduction corresponding to a 99% retained variance in their classification.

	Identification		Authentication	
Features	EO ACC	EC ACC	EO EER	EC EER
MCL	99.4 $\pm$ 0.4 %	98.8 $\pm$ 1.1 %	6.29 %	10.40 %
KFD	99.4 $\pm$ 0.7 %	98.9 $\pm$ 1.2 %	3.31 %	6.19 %
HM	99.5 $\pm$ 0.7 %	98.6 $\pm$ 1.0 %	3.50 %	6.12 %
PSD (beta and gamma)	97.4 $\pm$ 1.4 %	95.9 $\pm$ 1.6 %	28.40 %	31.80 %
AP (exponent and offset)	99.2 $\pm$ 0.6 %	99.2 $\pm$ 0.6 %	18.33 %	15.07 %
PLV† (beta and gamma)	100.0 %	100.0 %	1.35 %	2.63 %
COH† (beta and gamma)	100.0 %	99.8 $\pm$ 0.3 %	2.98 %	1.10 %
AR† (5th order)	97.9 $\pm$ 2.8 %	97.6 $\pm$ 1.9 %	6.11 %	5.33 %

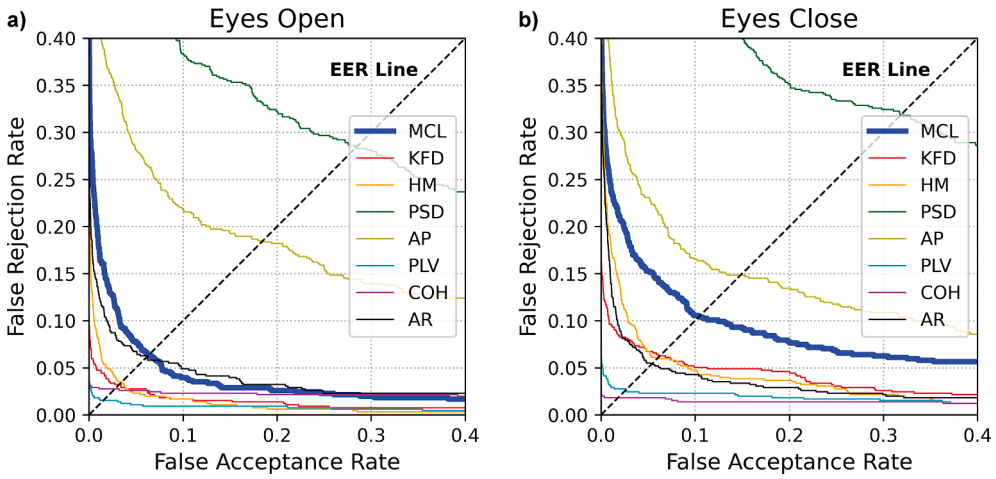


Fig. 4. Detection error trade-off (DET) curves for all features in the authentication task. Mean curve length (bold blue curve) lies below PSD and AP, which indicates that it performs better than them in the authentication. It lies above KFD, PLV and COH, indicating a poorer performance. The intersection point between each curve and the EER identity line is the equal error rate of the corresponding feature.

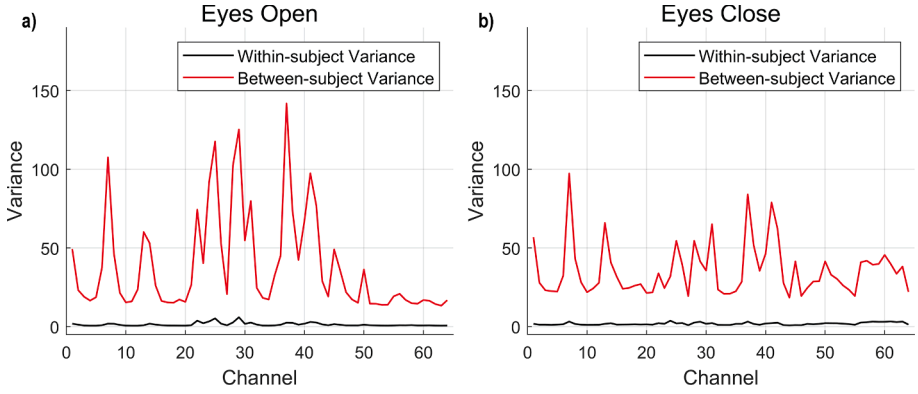


Fig. 5. Within-subject (black) and between-subject (red) variances of MCL for all channels in the (a) EO and (b) EC conditions. Subject 79 in EO and subjects 21 & 103 in EC were excluded as outliers.

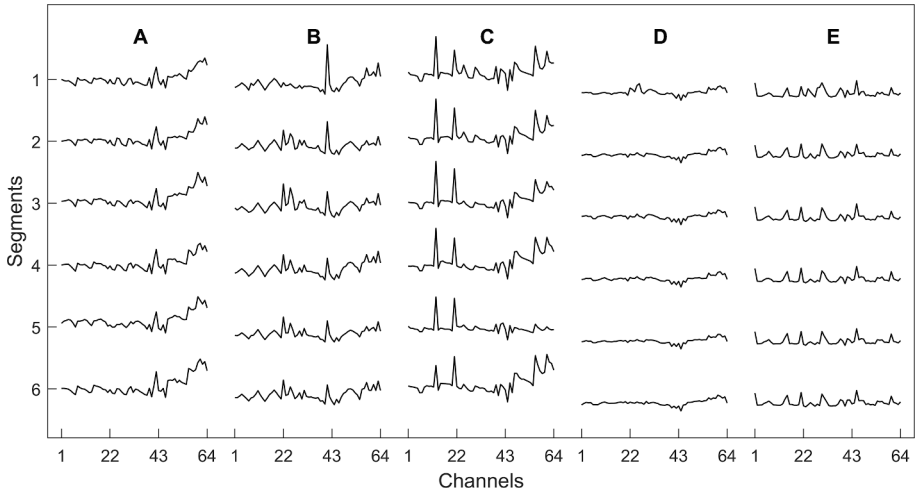


Fig. 6. MCL feature vectors (corresponding to features from 64 channels) of different temporal segments for five arbitrary subjects. The conspicuous similarity over time per each subject demonstrates MCL as a discriminative biometric feature.

3.2. Channel contributions to identification

We investigated channel contributions to the identification performance of MCL. A great overlap of signals between channels, hence redundancy of information, occurs for the scalp EEG because of the volume conduction effect. Each channel is expected to contribute to the performance depending also on its correlation with the other channels.

We adopted the so-called recursive feature elimination (RFE) method

for ranking channels based on their contribution. At each step, the highest identification performance for the remaining channels was marked after removing one channel. We observed that the identification accuracy for both conditions kept stable as long as the number of channels was about more than 20 (Fig. 7(a)). The highest accuracies (even a bit more than the one using the full channel set) were achieved with a minimum of 19 channels in EO (99.8%) and 27 channels in EC (99.7%).

Channels were ranked according to their elimination step, e.g., the lowest rank of 1 was assigned to the first removed channel, while the last one standing had the highest rank of 64. The topographical maps of the channel ranks shown in Fig. 7(b) hence indicate their contribution to the identification performance. The highest contributions mainly stemmed from the channels located at the parieto-occipital and temporal brain regions for both EO and EC conditions in accordance with recent literature localizing the spectral neural fingerprints in source level [14].

3.3. Effect of using mismatched covariance matrix

Retraining a Mahalanobis classifier consists in updating the covariance matrix and user templates when other users are needed to be either added to or removed from the database. We sought to investigate the degree of the importance of updating the covariance matrix in retraining. For this purpose, the robustness of the identification performance of MCL was questioned when a mismatched covariance matrix was employed for the classification.

We divided the 109 subjects randomly into two exclusive groups of unequal size: 20 (group I) and 89 (group II) subjects. Within each group, the subjects were classified in three ways: by computing the Mahalanobis distance (Eq. (6)) using the covariance matrix of i) the group itself, ii) the other group and iii) using no covariance matrix, i.e., the Euclidean distance. The average classification mean and standard deviation were obtained for a total of 500 random subject splits (Table 3).

We found out that using a mismatched covariance matrix still yielded a significantly higher performance in both groups and conditions than using plain Euclidean distance. The reduction in the identification performance was less in group I than group II, indicating that utilizing a larger population for estimating the covariance matrix mitigated the performance deterioration from the mismatch. Furthermore for the EC condition, a more robust estimation of the covariance matrix facilitated a better performance for group II than group I. This is despite that group II contains more subjects to classify, hence performing significantly worse than group I when plain Euclidean distance is used.

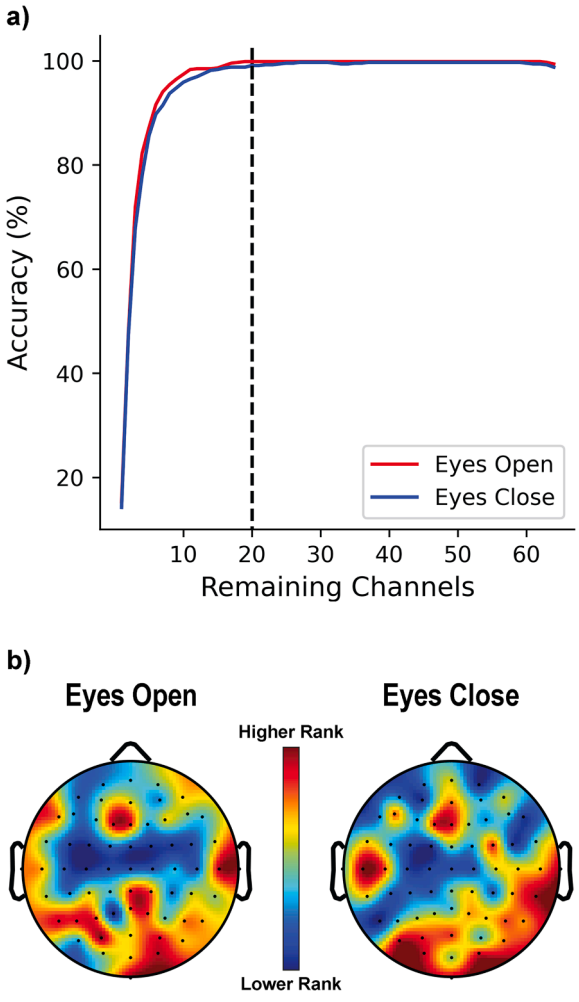


Fig. 7. (a) The maximum cross-validated identification accuracy via recursive feature elimination (RFE) was reached with a minimum of 19 channels (99.8%) and 27 channels (99.7%) in the EO and EC conditions, respectively. (b) Topographical views of the channel ranks for the EO and EC conditions. The most discriminating features are at the parieto-occipital and temporal regions.

4. Discussion

Nonlinear, higher order and complexity measures are notorious to have higher intricacy and to provide heavier computational costs as well as lesser interpretability of signals or systems underlying those signals. This is rather problematic for practical biomedical signal processing, particularly for the case of dense signals with high temporal resolution such as EEG, since efficiency is required for applicability to long data in short time and interpretability is desired to explain the mechanisms of the brain (in general, organs or biological systems). However for an unconventional metric, MCL is against the tide: it is simple, efficient, computationally cheap and descriptive.

Our study demonstrated MCL as a high performance while vastly efficient feature, when compared to the other well-established and highly discriminative features regularly employed in the state-of-the-art EEG-based biometric systems. In fact, our methodology is computationally so efficient, that it is fairly unconceivable to think of a metric simpler giving as much as high accuracy in identification of subjects from their cortical electrical activity. It should be noted that statistical measures of comparable simplicity such as power and variance do not provide the same level of individual distinction. Power spectrum and functional connectivity have provided the most preferred types of metrics to interpret electrical brain activity in literature. Our study showed that much simpler means like MCL also proves to be capable of distinguishing subject brains.

The performance of MCL was validated by having significantly

Table 3
Effect of using mismatched covariance matrices on MCL's identification performance. The values are the average obtained from 500 random iterations.

EO			
	Group I Covariance Matrix	Group II Covariance Matrix	No Covariance Matrix (Euclidean)
Group I (20 subjects)	99.5 ± 1.0 %	99.3 ± 1.1 %	97.3 ± 2.6 %
Group II (89 subjects)	98.1 ± 1.1 %	99.3 ± 0.7 %	96.0 ± 1.6 %
EC			
	Group I Covariance Matrix	Group II Covariance Matrix	No Covariance Matrix (Euclidean)
Group I (20 subjects)	98.5 ± 2.0 %	97.9 ± 2.0 %	94.1 ± 4.0 %
Group II (89 subjects)	96.6 ± 1.3 %	98.7 ± 1.1 %	89.8 ± 3.0 %

higher variance of between-subject than that of within-subject. The recognition performance of the state-of-the-art EEG-based biometric systems have already reached a satisfactory level of more than 95 %. Thus, the focus is gradually shifting towards addressing other obstacles that hinder EEG from becoming an acceptable biometric modality in practice. Despite the tremendous progress in the computational power in the last decades, efficiency and memory load still remain to be crucial for minimizing the delay in real-time biometric applications. Many methods presented in the literature can simply be too complex and slow to be practically applied in continuous authentication systems. In this regard, the remarkable speed of computation grants MCL a unique advantage to be embedded in real-world and real-time mobile biometric systems that exploit EEG brain signals endowed with high temporal resolution.

In addition to its computational efficiency, the implementation of MCL is also immensely easier, without the requirement of any specialized programming libraries. This is in contrast to the reliance on band-pass filtering, data transformation, curve fitting and third-party functions to perform the sophisticated calculations for many other measures such as PSD, AP, AR and functional connectivity matrices of PLV and COH. Moreover, there are no required user-specific input parameters in the computation of MCL. Whereas for most of the other features, there exist numerous estimation methods, basis functions and other filtering parameters to be selected, affecting the resultant performance. In addition to those parameters, there are multiple options for curve fitting for AP and a model order parameter for AR. In contrast, MCL is an accessible, easy-to-use feature for a naïve practitioner with no required consideration of intelligible inputs to the system, which should ideally reflect the characteristics of the underlying brain signals for accomplishing a proper estimation.

In our methodology, we performed a minimal preprocessing, only filtering out the high-frequency components of EEG signals beyond the conventional spectral bands mainly dominated by noise. Sophisticated artifact removal preprocessing may sometimes turn out to be counter-productive, as it may distort the signal and reduce practicality in real-life applications. In fact, some artifacts were even found to be useful for subject discrimination. For example, volume conduction effect [6,10] and muscle activity in high-beta and gamma [13] were shown to be contributing to the subject classification performance. For the classification, we took a template-based scheme and employed Mahalanobis distance as a measure of dissimilarity. Our results indicated that this simplistic approach provides satisfactory performance while containing a lower computational complexity. In particular, the modified version of Mahalanobis classifier was shown to be an effective and much faster alternative to deep learning and analytical feature selection approaches for high-dimensional features. Retraining the classifier, when adding or removing a user, solely requires updating the covariance matrix, which is considerably cheaper (especially when the feature vector is small) than a full retraining required for most other classifiers. Our analysis has revealed that as long as enough samples from many subjects are used for the estimation of the covariance matrix, regardless of being authorized users or not, the classification performance does not significantly deteriorate.

5. Conclusion

Neural fingerprints do not need to be obtained solely through conventional spectral band activities or functional connectivity metrics. In this direction, we proposed and demonstrated the effectiveness of MCL as a promising feature for EEG-based biometric identification and authentication. Despite its simplicity, MCL provides a remarkably unique characterization of individual brainwaves. Being a discriminative biometric feature in channel level together with its simplicity of formulation endows MCL with a powerful efficiency and capability desirable for real-time continuous brain biometrics. Our study indicates the potential usefulness of measures other than those derived from conventional EEG band activities and their functional interrelations and promotes further alternative analyses through exploiting the complexity of brainwaves.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Reza Yahyaei: Methodology, Software, Validation, Writing – original draft. **Tolga Esat Özkurt:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2022.103664.

References

[1] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (1) (2004) 4–20.
[2] Q. Gui, M.V. Ruiz-Blondet, S. Laszlo, Z. Jin, A survey on brain biometrics, ACM Comput. Surv. (CSUR) 51 (6) (2019) 1–38.
[3] A.J. Casson, Wearable EEG and beyond, Biomed. Eng. Lett. 9 (1) (2019) 53–71.
[4] J.C. Monsy, A.P. Vinod, EEG-based biometric identification using frequently occurring maximum power spectral features, IEEE Appl. Sig. Process. Conf. (2018) 249–252.
[5] K.P. Thomas, A.P. Vinod, EEG-based biometric authentication using gamma band power during rest state, Circ. Syst. Signal Process. 37 (1) (2018) 277–289.
[6] D.L. Rocca, P. Campisi, B. Vegso, P. Cserti, G. Kozmann, F. Babiloni, F.D.V. Fallani, Human brain distinctiveness based on EEG spectral coherence connectivity, IEEE Trans. Biomed. Eng. 61 (9) (2014) 2406–2412.
[7] K. Brigham, B.V.K. Vijaya Kumar, Subject identification from electroencephalogram (EEG) signals during imagined speech, IEEE Int. Conf. Biometr. Theory Appl. Syst. (2010) 1–8.
[8] D. La Rocca, P. Campisi, G. Scarano, On the repeatability of EEG features in a biometric recognition framework using a resting state protocol, Biosignals (2013) 419–428.
[9] S. Keshishzadeh, A. Fallah, S. Rashidi, Improved EEG based human authentication system on large dataset, Iran. Conf. Electr. Eng. (2016) 1165–1169.
[10] M. Fraschini, G. L. Marcialis, and L. Didaci, EEG-based personal identification: Comparison of different functional connectivity metrics, bioRxiv, 254557 (2018).
[11] A. Riera, A. Soria-Frisch, M. Caparrini, C. Grau, G. Ruffini, Unobtrusive biometric system based on electroencephalogram analysis, EURASIP J. Adv. Signal Process. 2008 (1) (2007).
[12] S. Zhao, K. Liu, X. Deng, EEG identification based on brain functional network and autoregressive model, IEEE Data Driven Control Learn. Syst. Conf. (2020) 474–479.
[13] M. Fraschini, A. Hillebrand, M. Demuru, L. Didaci, G.L. Marcialis, An EEG-based biometric system using eigenvector centrality in resting state brain networks, IEEE Signal Process Lett. 22 (6) (2015) 666–670.
[14] J. da Silva Castanheira, H.D. Orozco Perez, B. Misic, S. Baillet, Brief segments of neurophysiological activity enable individual differentiation, Nat. Commun. 12 (2021) 1.
[15] K.P. Thomas, A.P. Vinod, Biometric identification of persons using sample entropy features of EEG during rest state, IEEE Int. Conf. Syst. Man Cybernet. (2016) 3487–3492.
[16] Z. Mu, J. Hu, J. Min, EEG-based person authentication using a fuzzy entropy-related approach with two electrodes, Entropy 18 (2016) 432.
[17] N. A. Alzahab, M. Baldi, and L. Scalise, Efficient feature selection for electroencephalogram-based authentication, IEEE International Symposium on Medical Measurements and Applications, (2021) pp. 1–6.
[18] L.A. Moctezuma, M. Molinas, Event-related potential from EEG for a two-step identity authentication system, IEEE Int. Conf. Ind. Inform. 1 (2019) 392–399.
[19] K. Mukai and I. Nakanishi, Introduction of fractal dimension feature and reduction of calculation amount in person authentication using evoked EEG by ultrasound, IEEE Region 10 Conference (TENCON) (2020) pp. 567–572.

[20] M. Demuru, M. Fraschini, EEG fingerprinting: subject-specific signature based on the aperiodic component of power spectrum, Comput. Biol. Med. 120 (2020), 103748.
[21] T. Higuchi, Relationship between the fractal dimension and the power law index for a time series: a numerical investigation, Phys. D 46 (1990) 254.
[22] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless, Line length: An efficient feature for seizure onset detection, 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 2 (2001) pp. 1707–1710.
[23] L. Guo, D. Rivero, J. Dorado, J.R. Rabuñal, A. Pazos, Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks, J. Neurosci. Methods 191 (1) (2010) 101–109.
[24] L. Logesparan, A.J. Casson, E. Rodriguez-Villegas, Optimal features for online seizure detection, Med. Biol. Eng. Compu. 50 (7) (2012) 659–669.
[25] T. E. Özkurt, M. Sun, T. Akgül, and R. J. Sclabassi, Optimal feature selection for seizure detection: A subspace based approach, International Conference of the IEEE Engineering in Medicine and Biology Society (2006) pp. 2134–2137.
[26] J. Chen, J. Benesty, Y.A. Huang, Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments, EURASIP J. Adv. Signal Process. 2005 (2005) 1.
[27] L. Hui, B.-q. Dai, and L.Wei, A pitch detection algorithm based on AMDF and ACF, in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 1 (IEEE, 2006) pp. I–I.
[28] T. Akgül, M. Sun, R. Sclabassi, A.E. Çetin, Characterization of sleep spindles using higher order statistics and spectra, IEEE Trans. Biomed. Eng. 47 (2000) 997.
[29] T.E. Özkurt, H. Akram, L. Zrinzo, P. Limousin, T. Foltynie, A. Oswal, V. Litvak, Identification of nonlinear features in cortical and subcortical signals of Parkinson's disease patients via a novel efficient measure, Neuroimage 223 (2020) 117356.
[30] M. Wang, J. Hu, H.A. Abbass, Brainprint: EEG biometric identification based on analyzing brain connectivity graphs, Pattern Recogn. 105 (2020), 107381.
[31] G. E. Hine, E. Maiorana, and P. Campisi, Resting-state EEG: A study on its non-stationarity for biometric applications, International Conference of the Biometrics Special Interest Group (2017) pp. 1–5.
[32] J.C. Monsy, A.P. Vinod, EEG-based biometric identification using frequency-weighted power feature, IET Biom. 9 (6) (2020) 251–258.
[33] T. Nakamura, V. Goverdovsky, D.P. Mandic, In-ear EEG biometrics for feasible and readily collectable real-world person authentication, IEEE Trans. Inf. Forensics Secur. 13 (3) (2018) 648–661.
[34] T. Schons, G. J. P. Moreira, P. H. L. Silva, V. N. Coelho, and E. J. S. Luz, Convolutional network for EEG-based biometric, Springer Iberoamerican Congress on Pattern Recognition (2017) pp. 601–608.
[35] L. Ma, J.W. Minett, T. Blu, W.-S.-Y. Wang, Resting state EEG-based biometrics for individual identification using convolutional neural networks, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 2848–2851.
[36] Y. Sun, F.-W. Lo, B. Lo, EEG-based user identification system using 1D-convolutional long short-term memory neural networks, Expert Syst. Appl. 125 (2019) 259–267.
[37] E. Maiorana, Deep learning for EEG-based biometric recognition, Neurocomputing 410 (2020) 374–386.
[38] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, circulation 101, e215 (2000).
[39] B. Singh, S. Mishra, U.S. Tiwary, EEG based biometric identification with reduced number of channels, in: IEEE 17th International Conference on Advanced Communication Technology, 2015, pp. 687–691.
[40] B.B. Das, P. Kumar, D. Kar, S.K. Ram, K.S. Babu, R.K. Mohapatra, A spatio-temporal model for EEG-based person identification, Multimedia Tools Appl. 78 (19) (2019) 28157–28177.
[41] M. DelPozo-Banos, C.M. Travieso, J.R. Ticay-Rivas, J.B. Alonso, M.K. Dutta, A. Singh, Real cepstrums on electroencephalogram biometric identification, IEEE Int. Conf. Contem. Comput. (2015) 54–58.
[42] M. Wang, H. El-Fiqi, J. Hu, H.A. Abbass, Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states, IEEE Trans. Inf. Forensics Secur. 14 (12) (2019) 3259–3272.
[43] M. Wang, J. Hu, H.A. Abbass, Stable EEG biometrics using convolutional neural networks and functional connectivity, Aust. J. Intell. Inf. Process. Syst. 15 (2019) 19.
[44] S.u. Yang, F. Deravi, S. Hoque, Task sensitivity in EEG biometric recognition, Pattern Anal. Appl. 21 (1) (2018) 105–117.
[45] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Phys. D 31 (2) (1988) 277–283.
[46] L.A. Moctezuma, M. Molinas, Classification of lowdensity eeg for epileptic seizures by energy and fractal features based on emd, J. Biomed. Res. 34 (2020) 180.
[47] J.E. Jacob, G.K. Nair, A. Cherian, T. Iype, Application of fractal dimension for eeg based diagnosis of encephalopathy, Analog Integr. Circ. Sig. Process 100 (2) (2019) 429–436.
[48] D. Parbat, M. Chakraborty, A novel methodology to study the cognitive load induced eeg complexity changes: chaos, fractal and entropy based approach, Biomed. Signal Process. Control 64 (2021), 102277.
[49] K.J. Friston, Functional and effective connectivity in neuroimaging: a synthesis, Hum. Brain Mapp. 2 (1994) 56.
[50] M.J. Katz, Fractals and the analysis of waveforms, Comput. Biol. Med. 18 (3) (1988) 145–156.
[51] J. Monge-Álvarez, Higuchi and Katz fractal dimension measures, https://www.mathworks.com/matlabcentral/fileexchange/50290-higuchi-and-katz-fractal-dimension-measures, Last accessed: 26.12.2021 (2021), MATLAB Central File Exchange.
[52] R. Vallat and J. V. D. Donckt, AntroPy, https://github.com/raphaelvallat/antropy/blob/master/antropy/fractal.py, Last accessed: 26.12.2021 (2021), GitHub.
[53] E. Shamsi, M.A. Ahmadi-Pajouh, T.S. Ala, Higuchi fractal dimension: an efficient approach to detection of brain entrainment to theta binaural beats, Biomed. Signal Process. Control 68 (2021), 102580.
[54] I. Wijayanto, R. Hartanto, H.A. Nugroho, Multidistance fluctuation based dispersion fractal for epileptic seizure detection in eeg signal, Biomed. Signal Process. Control 69 (2021) 102938.
[55] F. Scarpa, M. Rubega, M. Zanon, F. Finotello, A.-S. Sejling, G. Sparacino, Hypoglycemia-induced eeg complexity changes in type 1 diabetes assessed by fractal analysis algorithm, Biomed. Signal Process. Control 38 (2017) 168–173.
[56] R. Esteller, G. Vachtsevanos, J. Echauz, B. Litt, A comparison of waveform fractal dimension algorithms, IEEE Trans. Circuits Syst. I: Fundamental Theory Appl. 48 (2) (2001) 177–183.
[57] T. Donoghue, M. Haller, E.J. Peterson, P. Varma, P. Sebastian, R. Gao, T. Noto, A. H. Lara, J.D. Wallis, R.T. Knight, A. Shestyuk, B. Voytek, Parameterizing neural power spectra into periodic and aperiodic components, Nat. Neurosci. 23 (12) (2020) 1655–1665.
[58] J. Walters-Williams and Y. Li, Comparative study of distance functions for nearest neighbors, Advanced techniques in computing sciences and software engineering (2010) pp. 79–84.
[59] A. Kessy, A. Lewin, K. Strimmer, Optimal whitening and decorrelation, Am. Statist. 72 (4) (2018) 309–314.