Biomedical Physics & Engineering Express

Deep-learning online EEG decoding brain-computer interface using error-related potentials recorded with a consumer-grade headset

Dorina-Marcela Ancau[1], **Mircea Ancau**[2] and **Mihai Ancau**[3,*]

1   Technical College for Transportation 'Transylvania', Cluj-Napoca, Romania
2   Department of Industrial Engineering, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
3   Department of Neurology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
*   Author to whom any correspondence should be addressed.

E-mail: dorina.ancau@colegiuldetransporturi-cluj.ro, mircea.ancau@tcm.utcluj.ro and mihai.ancau@tum.de

Abstract

Objective. Brain-computer interfaces (BCIs) allow subjects with sensorimotor disability to interact with the environment. Non-invasive BCIs relying on EEG signals such as event-related potentials (ERPs) have been established as a reliable compromise between spatio-temporal resolution and patient impact, but limitations due to portability and versatility preclude their broad application. Here we describe a deep-learning augmented error-related potential (ErrP) discriminating BCI using a consumer-grade portable headset EEG, the Emotiv EPOC$^+$. *Approach.* We recorded and discriminated ErrPs offline and online from 14 subjects during a visual feedback task. *Main results:* We achieved online discrimination accuracies of up to 81%, comparable to those obtained with professional 32/64-channel EEG devices via deep-learning using either a generative-adversarial network or an intrinsic-mode function augmentation of the training data and minimalistic computing resources. *Significance.* Our BCI model has the potential of expanding the spectrum of BCIs to more portable, artificial intelligence-enhanced, efficient interfaces accelerating the routine deployment of these devices outside the controlled environment of a scientific laboratory.

1. Introduction

A Brain-Computer Interface (BCI) is a communication and control system optimized for decoding bioelectric cortical neuronal activity and transforming it into a control signal for a specific clinical application. For people with neuromuscular disabilities or neurodegenerative diseases with serious motoric deficits, it is an alternative to natural communication or neuromuscular pathways, acting in a sense as an artificial bypass.

Evoked potentials or event-related potentials (ERPs) can be utilized to guide BCIs. ERPs are any stereotypical responses of the brain to a stimulus, be it external or internal, sensory, cognitive or motoric, that can be measured for practical purposes via electroencephalography (EEG) and originate from infinitesimal variations of the electric field potential due to the extracellular summation of synchronous postsynaptic currents of many neurons from the surface of the cerebral cortex upon repetition of the same eliciting paradigm [1]. The recorded brain activity is channelled through several data processing steps in order to decode the user's intended action, which is then used to encode a command to a device that can be observed by the user as feedback, or which performs an intended and trainable action [2].

To date, the paradigm most non-invasive BCIs have focused on has been the visually evoked potential and, in particular, its most prominent positive deflection, the P300 component (or, in short, P3), named by convention for its polarity and relative latency to stimulus onset, with the P3 speller being one of the most commonly used applications in BCI systems [3]. However, these potentials are usually recorded using high-resolution, professional EEG equipment, which, besides the large number of electrodes and therefore a cumbersome attachment to the scalp, requires a high acquisition cost.

The past few years have brought about a quantum leap in the development of low-cost, affordable and portable EEG devices, extending the application sphere of BCIs 'from bench to bedside' [4]. One such device is the Emotiv EPOC$^+$, a portable, multichannel interface. It was designed, initially, for 'mind hacking', but proved its potential in the acquisition and analysis of EEG signals in preclinical settings as well, mostly focusing on the P3 component, which has been extensively studied and was therefore easier to establish [5–8].

In this regard, error-related potentials (ErrPs) represent a particular case of endogenous ERPs, which result, in experimental settings, when another subject or (virtual) machine is being interacted with and interprets the main user's intention in an erroneous way. The structure of one such stereotypical signal, typically originating from cortical electrical dipoles in the posterior medial frontal cortex (pMFC) [9] or in the anterior cingulate cortex (ACC) [10], both corresponding to the Cz electrode, shows a positive component some 200 ms after stimulus onset and a further negative deflection around 250 ms followed by another ample positive component at ca. 320 ms, in short, the N_{200}/P_{300} complex [11]. Conveniently, ErrPs by nature encode a binary/boolean response (true or false, correct or erroneous) to a choice task. In practice, an agent equipped with an ErrP-BCI (i.e., a patient suffering from tetraplegia) could thus, by mentally marking the wrong choice and generating an ErrP, select the correct one from a range of two possibilities offered by another agent (i.e., a robot able to perform daily tasks). Since any n-fold multiple-choice task can be subdivided into (n–1) consecutive 2-fold or binary choice tasks (by recursively leaving out one of the remaining choices and pooling all the others into one single choice that gets subdivided in the next step), one can theoretically conceive a method of deciphering a complex intention via an algorithm able to decompose a complex task in multiple successive binary subtasks, which a subject could then express his/her choice for, by the issuing of ErrPs alone.

One of the practical challenges in the fields of machine learning and of BCI training is represented by the difficulty of training reliable and generalizable classifier algorithms with limited subject/patient data [12]. Due to the dependency on human subject data, gathering a sufficiently wide and varied training set for precise online data classification is more challenging even than in other fields of machine learning, such as computer vision. We inferred that either unsupervised learning deep convolutional generative adversarial networks (DCGANs) [13] or time-frequency-representations, such as intrinsic mode functions (IMFs) [14] could be employed to circumvent this conundrum, by augmenting the ErrP dataset after discretization in epochs, in a similar way that image datasets are augmented in computer vision [13, 15].

Another challenge in the field of BCIs relates to improving algorithm classification performance for practical use in real time applications, particularly in aspects such as classification accuracy, speed and generalization capacity across multiple subjects [16]. Historically, gaussian classifiers [17], Bayesian filters [18], logistic regression, k-nearest neighbours [19], linear/ quadratic discriminant analysis and support vector machines or combinations of the latter [20] have been employed after heterogenous data pre-processing protocols in the classification of ErrPs, achieving classification performances between 70% and 80% with professional EEG recording equipment [16]. With the burgeoning expansion of machine learning applications by solving the problem of vanishing gradients [21] in fields such as computer-aided vision, image recognition, speech and text analysis, interest has grown for implementing these technologies to the field of EEG signal processing [16].

We surmised therefore that a supervised machine learning algorithm based on a Long-Short Term Memory Neural Network—a special case of Recurrent Neural Network (RNN) [22]—could be deployed to learn features from pre-classified ErrP data epochs and achieve higher online classification accuracies than traditional techniques when combined with an artificially augmented data set.

2. Methods

A total of 14 healthy subjects, 10 male and 4 female, aged between 21 and 61 years (mean 30.2 years) and with no history of neurological or psychic disorders participated in this study. They all expressed informed consent for voluntary participation in conformity with the local ethical regulations.

2.1. Experimental paradigm

The experimental paradigm was optimized to elicit a visual interaction feedback ErrP, characterized by a feedback-related negativity (FRN) component appearing between 200 and 300 ms after feedback/stimulus onset [11]. Each participant was instructed to perform a multiple-choice reaction human-computer interaction task [23, 24] created in *Psychtoolbox* [25], an extension of the programming language *Matlab* (version R2012b, *Mathworks Inc.*). The subject was required to direct the movements of a virtual device (a blue square on a black background) towards the indicated target, which was represented by a green square located in one of three possible positions, by quickly pressing an appropriate key, while in each of the other two positions a distracting red square was placed. (figure 1). A correct trial was signified by the virtual device (the blue square) moving towards the target (the green square), while any movement towards one of the distractors (the red squares) was deemed an incorrect trial. The participants were

Figure 1. Experimental paradigm and data acquisition pipeline.
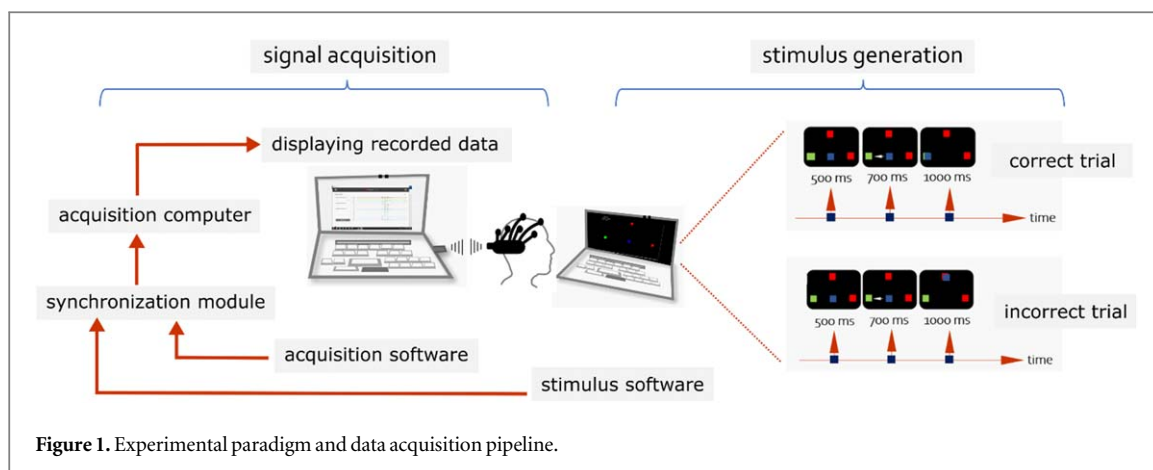
comfortably seated 1 m away from a computer screen displaying all and only the information related to the experiments. They were allowed to freely move their eyes and to blink, while other muscular movements were restricted to resting periods.

Upon key press, signifying one trial, the virtual device would in 25% of the trials move towards a non-target stimulus, even if the user response were correct, the disjunction between the intention and the feedback potentially generating the FRN. The protocol was based on previous findings [10]. Each subject conducted one session comprising four blocks of 30, 60 or 100 trials each, lasting between 5.5 and 14 min over the course of one day. The initial 20% of trials in each block were designated as training runs and served the purpose of training the neural network classifier.

2.2. Headset

For our study, we employed the consumer-grade Emotiv EPOC$^+$ EEG headset with 14 (+2) channels corresponding to the standard EEG electrode locations: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 (locations P3 and P4 represent CMS and DRL electrodes, mass and reference potential, respectively) [26]. It represents a lightweight EEG headset with a single analog-to-digital converter with a maximum sampling frequency of 256 Hz, performing conversion and pre-processing as well as wireless transmission to a USB receiver, and therefore allowing EEG recording under naturalistic circumstances [17]. It has already been evaluated, along with other consumer-grade headsets, among them NeuroSky's MindWave, for the P3 speller paradigm, generating somatosensory, steady-state visual evoked potentials or basic image or emotion recognition [27–30]. Recorded and pre-processed signals are passed by the headset itself through a bandpass filter of 0.2–45 Hz. The fidelity of the recordings depends heavily on the quality of the electrode contacts, whose state is continuously monitored and fed back in the form of an impedance index to the user via the acquisition software, Emotiv PRO. One downside to the compactness of the headset with regards to our paradigm was

the lack of a Cz electrode, which would have been especially suitable for recording ErrPs due to its proximity to the ACC, so that our study focused on its two nearest available electrodes, FC5 and FC6.

2.3. Signal acquisition

Once electrode contact was established, a consumer-grade laptop (ASUS ZenBook UX310) was deployed running the Emotiv PRO software, from which the 14-channel data stream in the 0.2–45 Hz frequency range was redirected to a custom-designed driver written in Python (Python Software Foundation. Python Language Reference, version 3.7; available at http://www.python.org) extracting data via the Emotiv Cortex 1.9 API (available at https://emotiv.gitbook.io/cortex-api/), allowing them to communicate with other programming languages and hardware (figure 1). Accessing raw EEG data requires licensing from Emotiv and, as of Cortex 1.9 API, a subscription-based authentication token inherently linked to the Emotiv user identification code. The Cortex 1.9 API could be communicated with via WebSocket Secure protocol server using a JSON-RPC (Java Script Object Notation - Remote Procedure Call). For interfacing with the EEG stream and online signal display and classification we designed a virtual instrument using LabVIEW (Laboratory Virtual Instrument Engineering Workbench) [31], a complex visual programming environment optimized for quick translation to industrial platforms. From the front panel, real-time data flow on the two relevant electrodes (FC5 and FC6) as well as the numerical values of EEG amplitudes were displayed. In the background, the instrument employed a *System Exec VI* to launch and execute a custom *Python* script that fed continuous data from the headset into an array through a *Spreadsheet* string block. As LabVIEW is optimized for designing electronic circuit pipelines, signal analysis and classification was achieved with Matlab by the plugins EEGLAB [32] and ERPLAB [33], while neural network design and training was performed in Matlab.

2.4. Signal processing

Data preprocessing in EEGLAB involved reading the signal stream as well as event timing information, together with electrode positions, followed by filtering (0.2–12 Hz band-pass filter with Hamming window and a transition bandwidth of 4 Hz), relevant electrode selection (FC5 and FC6), discretization into epochs (from 0 until +900 ms after stimulus onset), application of Independent Component Analysis (ICA) and hybrid artifact rejection methods (improbable data epochs of amplitudes over 25 μV, abnormal data trends, abnormal frequency spectrum amplitudes of over ± 25 dB, values higher/lower than 8 standard deviations of the data amplitude distribution, as well as visual signal inspection) followed by the calculation of grand averages across all of the epochs pertaining to one subject and the same paradigm.

2.5. Signal data augmentation

Preclinically or clinically recorded EEG data inherently suffers from low signal-to-noise and samples-to-feature ratios, imbalanced datasets (in the case of rare target events) as well as non-stationarity [34], thereby having poor inter- and intraindividual generalization potential and significantly reducing effectiveness of deep learning strategies [35]. Not unlike deep learning revolutionized computer vision, we supposed that data augmentation approaches might lead to increased classifier accuracy, stability and, in the case of artificial neural network classification, reduced overfitting [35]. The aim would be to artificially generate new data samples increasing the quantity and diversity of the original data based on existing training data [34], either by Empirical Mode Decomposition (EMD)—in the one case via Intrinsic Mode Functions (IMFs)—or by using a hidden model to create artificial data with a similar distribution to the real data—in the other case via Deep Convolutional Generative Adversarial Networks (DCGANs).

2.6. Deep convolutional generative adversarial networks

Encouraged by the classification improvement after image dataset augmentation with DCGANs in computer vision research [13], DCGANs, a more stable subset of Generative Adversarial Networks (GANs) based on Convolutional Neural Networks (CNNs), addressing the problem of meaningless output during network training [34], have begun to be employed in generating and labelling artificial EEG signal for classifier training [12]. To simplify, GANs consist of a generator G generating artificial data from a tensor initialized with Gaussian noise of a particular dimension and trying to confuse the discriminator network D, which continuously learns and revises its discerning ability, into classifying generated data as original data, as in a minimax game until the loss function L [36] is minimized, the system reaching a Nash equilibrium

[34] (here x is the real data distribution, while y represents the generator output):

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log D(x_i)$$
$$\qquad -\frac{1}{N}\sum_{i=1}^{N}(1 - y_i)\log(1 - D(x_i)) \qquad (1)$$

In analogy with image processing, G creates filters (or kernels, not to be confused with the bandpass filters from the signal processing pipeline) based on the CNN learning pathway and ensures by calibration through D that they learn relevant features for the recognition of the target data. Theoretically, in the Nash equilibrium state, the distributions of the original and the artificial data coincide, because D cannot distinguish whether data from G is artificial or real [34]. In our case, we used a magnification factor of 2, i.e., 50% of the data for training was original, while the rest was generated with DCGANs. On the one hand, our G converted a Gaussian noise tensor of size $1 \times 1 \times 100$ to a $160 \times 1 \times 511$ tensor using a *project and reshape* layer, which then gradually rescaled the output using a series of transposed convolution layers with 1×5 filters with a decreasing number of filters for each layer, interspersed batch normalization layers, in order to avoid internal covariate shifting, and *ReLU* layers. On the other hand, D takes the output of G, a $231 \times 1 \times 1$ tensor (231 being the number of samples in an epoch at a sampling rate of 256 Hz) and returns a scalar prediction score using a series of convolution layers with batch normalization and leaky *ReLU* layers (figures 2(a)–(c)). The numerical relationships between the dimensions of the convolutional layers were determined mathematically using a system of linear first-order equations based on the formula:

$$N_{k+1} = \frac{N_k + 2p_k - f_k + 1}{s_k} \qquad (2)$$

where N_k represents the number of activations (neurons, units) of convolutional layer k, p_k—the padding of the matrix of convolutional layer k (padding means the process of adding zeros to the input matrix symmetrically), f_k—the size of convolutional filter k (set of learnable weights which are learned using the backpropagation algorithm) and s_k—the stride of convolutional layer k (the number of units/neurons shifted over the input matrix). Given the initial number of samples in an epoch (231, see above) and the number of convolutional layers in the network—which matches those frequently employed in the literature—one can determine for every layer of the network the parameters N_k, p_k, f_k, s_k recursively, so that the apparently arbitrary sizes of the layers become mathematically determined and optimal in the context of the current task.

2.7. Intrinsic mode functions

Another approach for artificially expanding a set of EEG data epochs involves the calculation of the harmonics of each epoch and recombining them in
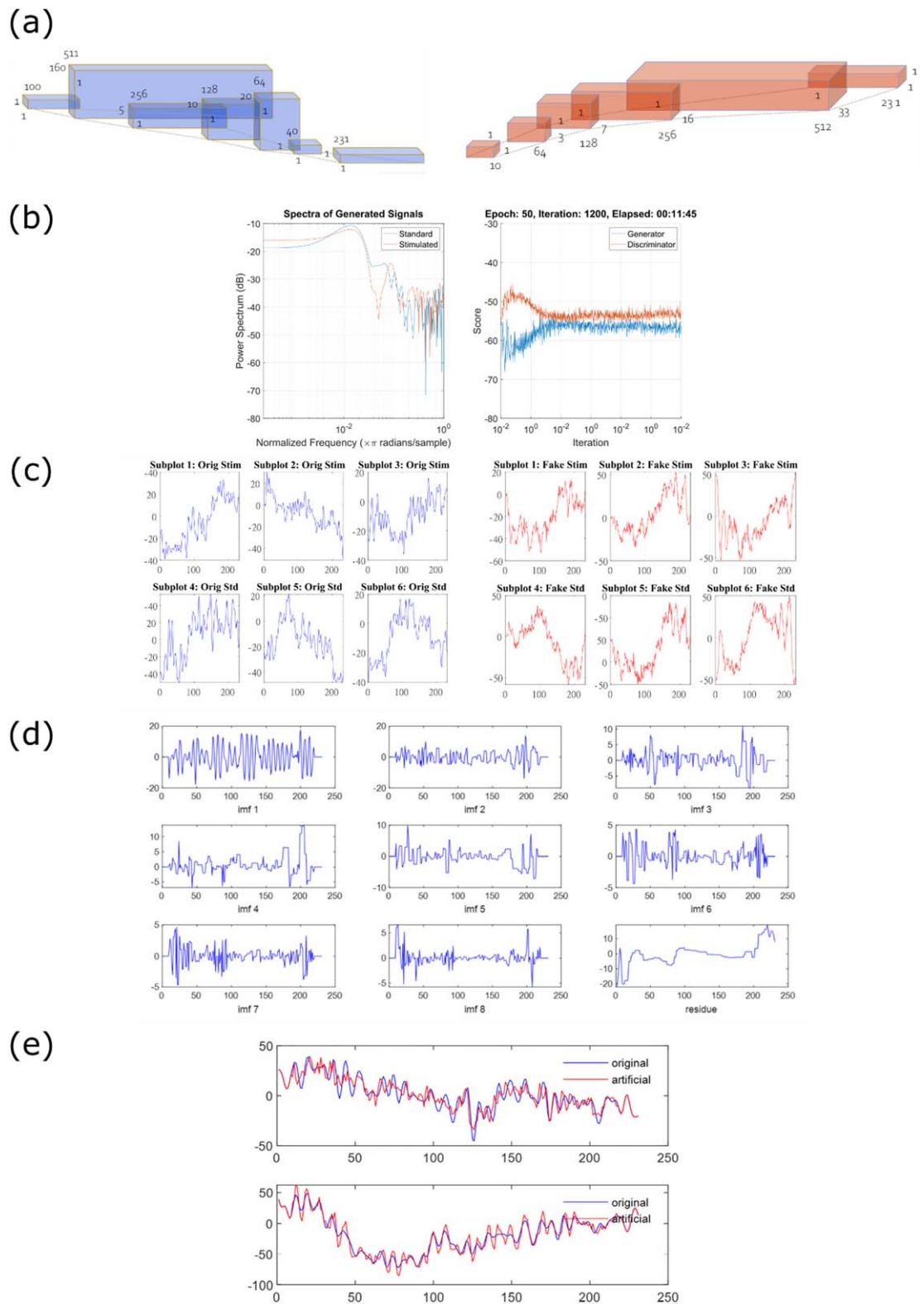
Figure 2. Data augmentation methods. (a) Deep convolutional generative adversarial network (DCGAN). On the left, in blue, the generator network, providing a 231 × 1 × 1 input to the discriminator network, on the right, in red. (b) Example of a training run of the DCGAN. Left: Power spectrum of the original (blue) against the artificial (red) signal. Right: Training plot of the scores of the generator (blue) and discriminator (red) networks. (c) Example plots from the training run in (b), displaying the first 6 original epochs (blue, left) and the first 6 artificial epochs (red, left) generated by the DCGAN. (d) Example of the first 8 IMFs and the residue generated from an error trial. Artificial epochs were generated by combining IMFs of corresponding orders between original epochs. (e) Examples of one standard (above, blue) and one error trial epoch (below, blue) along with the corresponding IMF-generated artificial epochs (above and below, red).

order to generate artificial ones. Since EEG signals are non-periodic and non-stationary, simply applying a Fourier transform to the signal would lead to a meaningless frequency spectrum [37]. However, even in these circumstances EEG signals can be decomposed into a finite number of nonlinear harmonics of the initial signal, called IMFs, by applying the EMD method [14]. These IMFs fulfil two conditions, namely: first, the number of endpoints must be equal to the number of zero crossings, or differ by at most one; secondly, the average value of the envelopes defined by the maximum and minimum endpoints, respectively, must be zero. The process of generating IMFs from an EEG signal *x(t)* consists theoretically of the following steps:

(a) Find the maximum and minimum points of the function *x(t)*. If the number of maximum or minimum points is sufficient to be able to calculate their envelopes, then proceed to step b). Here, each type of interpolation (linear, polynomial, Bézier, spline etc) requires a minimum number of points. The envelope of the maxima and minima is calculated in this case by cubic spline interpolation, for which a minimum number of 8 points are necessary.

(b) Find, by interpolation, the upper and lower envelope of the end points, respectively. Depending on the geometry of the signal, a point of maximum amplitude of the signal can lie below zero, just as a point of minimum amplitude can lie above zero. The upper and lower envelopes designate the envelopes of all maxima or all minima of the signal.

(c) Calculate, for each point *t*, the *mean(t)* of the two envelopes.

(d) Calculate the difference *diff(t) = x(t) − mean(t)*.

(e) If *diff(t)* meets the two conditions to be able to constitute an IMF, then *imf_i = diff(t)*.

(f) Subtract *diff(t)* from the initial signal and at the next step *x(t) ← x(t) - diff(t)*, at which point restart with the new *x(t)* from step (a).

(g) If *diff(t)* does not meet the two conditions mentioned above, then *x(t) ← diff(t)* and then return to step (a).

(h) If the number of maximum and minimum points, respectively, are not sufficient for the calculation of the envelopes, then the process stops, and the residue is *res(t) ← diff(t)*.

Once the IMFs have been determined, the initial signal can be restored by simply summing up the IMFs together with the residue:

$$x(t) = \sum_{k=1}^{n} imf_k(t) + res(t) \qquad (3)$$

The total number of generatable IMFs per sample varies with the structure of the EEG signal. In our study, we extracted 12 successive IMFs from each epoch, the remaining signal being designated as residue. To generate artificial epochs, we recombined IMFs from the same trial (and subject) (figures 2(d)–(e)). Each recording comprised 60 epochs, out of which 15 contained an ErrP stimulus (*stim*) while 45 did not (*std*). We proceeded as in the following example in the case of *stim* epochs: imf_1 from epoch 1 replaced imf_1 from the other epochs thus helping generate 14 artificial *stim* epochs. Then imf_1 from epoch 2 replaced imf_1 from the other epochs and so on. In this way, 196 artificial *stim* epochs were created. The analogous procedure was performed with imf_2, thus generating other 196 artificial epochs and so on. The same was then performed with *std* epochs.

2.8. Online signal classification using long short-term memory networks

Lately, implementation of deep neural networks (DNNs) has yielded promising results in the classification of linguistic, imagistic and acoustic data [34, 38]. DNNs minimize redundant information bias and improve general classifier accuracy, but especially in EEG signal decoding they are impaired by the small number of available samples and suffer from poor generalizability [34]. LSTMs constitute an advancement of RNNs in recognition of the issue of vanishing/exploding gradients upon learning long data sequences [22]. They are particularly suitable for studying word sequence and time-series data, by learning long-term dependencies between data units. In our study, we adapted a bidirectional LSTM network initially used for ECG signal classification [39], consisting of a bidirectional LSTM layer with an output size of 100, which is fed into a fully connected layer of size 2, followed by a *softmax* layer and finally a classification layer.

2.9. Online classification using LDA and QDA

As a benchmark for classical supervised machine learning classification algorithms, we employed either a Linear (LDA) or Quadratic Discriminant Analysis (QDA), also as a means to compare the methods. In LDA, the distance between the averages of the two classes to be discriminated is maximized, while the variability within each class is minimized, according to Fisher's criterion, expressed as:

$$J(w) = \frac{|\widetilde{\mu}_1 - \widetilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}; \qquad (4)$$

where $\widetilde{\mu}_1$, $\widetilde{\mu}_2$ represent the averages of the projections of the classes along the direction *w*, and $\tilde{s}_1^2$, $\tilde{s}_2^2$ represent the measure of the variability within the two classes.

LDA seeks a vector *w* for which *J(w)* is maximal. In other words, *J(w)* has a maximum value when the distance between the averages of the class projections is maximum and the measure of variability within the classes is minimal. Two types of classifiers were generated in *Matlab*, a linear and a quadratic one. The classes to be discriminated against were the sets of standard epochs and stimulated epochs from each recording. The 231 potential values within an epoch were the classifier features and corresponded to an interval of 900 milliseconds (sampled at 256 Hz). For the relevant features of the epoch, according to the type of stimulus, the interval from millisecond 100 to 500 after the stimulus was considered, since relevant ErrP components (such as N_2 or P_3) are known to appear in this interval.

The classifier (LDA or QDA) would then attempt to express one dependent variable (signal) as a linear or quadratic combination of the independent variable (time). As in other cases of machine learning (such as principal component analysis or neural network models), the exact features of the variables (signal) selected for best discrimination between classes by the algorithm can only poorly be explained or predicted intuitively in human terms, making the inner workings of such algorithms obscure.

To reduce the number of data points in an epoch, the number of values within the interval 100 and 500 ms was divided into subintervals of 50 milliseconds with 50% overlap and the average value within it was considered to be the feature of each group. I.e., group 1 ranged from 100 ms to 150 ms, group 2 from 125 ms to 175 ms, and so on. This reduced the number of features with which the algorithm operated from 103 per epoch (spanning 400 ms from 100 to 500 ms at 256 Hz sampling rate) to effectively 8 features (8 intervals of 50 ms spanning the 400 ms from 100 to 500 ms after stimulus onset).

2.10. Statistical analysis

We did not employ a statistical power analysis in order to determine the minimal number of subjects and trials, but our sample sizes match those generally employed in the field. Statistical analysis was performed using R (https://www.R-project.org/, version 4.1.0) and R Studio (version 1.4.1717) software. Data were tested for normality and equality of variance using the Shapiro-Wilk test and, respectively, the Bartlett test.

If the data passed normality and homoscedasticity criteria, we performed one-way repeated-measures ANOVA tests with Bonferroni adjustments for multiple comparisons. If they did not, then we performed the Kruskal-Wallis test with Dunn's multiple comparisons test. All statistical tests were two-sided and performed to a significance level of $p <= 0.05$.

3. Results

3.1. Offline ErrP detection

As an internal control, we started by verifying that ErrPs can be validly detected with Emotiv EPOC$^+$ in the first place, by analysing offline recordings from just 3 of the subjects, which performed the same feedback ErrP generating task over the course of multiple sessions. The characterization of the shape and timing of the potentials and their components was achieved through the computation of time-locked grand average potentials for the correct (standard, *std*) and error (stimulated, *stim*) trials in channels FC5 and FC6 [24]. A standard topographic interpolation of the potentials was calculated every 100 ms from timepoint $t = 0$ (feedback stimulus onset) to 500 ms after stimulus onset. Single-trial potentials in both trial conditions were plotted as a colour-encoded image (figure 3(b)) with a smoothing window of 50 trials, as in [24].

In particular, we observed that grand averages ErrP time courses from epochs following ErrP eliciting stimuli (*stim*) exhibited an FRN/N2 component between 300 and 450 ms on frontocentral electrodes FC5 and FC6 following stimulus onset and a P3 component between 450 and 600 ms, which have been consistently reported in ErrP eliciting assays [11, 40]. The higher latency than in previous studies, where the FRN generally appeared between 200 and 350 ms on electrodes Cz and FCz following stimulus onset [11], could be, in our opinion, attributed to different electrode positioning due to headset-inherent constraints. Whereas professional EEG caps allow recordings from the electrode position Cz, in the case of Emotiv EPOC+, which was not primarily designed for recording ErrPs, this position is not represented, therefore the closest available electrodes were selected, FC5 and FC6.

Grand average ErrPs across subjects revealed on both FC5 and FC6 statistically significant (1-way ANOVA) negative signal components in the 200 to 450 ms range on the *stim* condition (figure 3(d)). Scalp 2D and 3D mapping hinted towards a frontocentral localization of the component origins (figure 3(a)), while power spectrum analysis revealed largely identical frequency representations with a slight bias towards lower frequencies (2 to 3.5 Hz) on electrode FC6 (figure 3(c)). Altogether the data were compatible with valid recordings of feedback ErrPs, so that we proceeded to the live recording and classification scenario.

3.2. Online ErrP classification

Online classification of ErrPs was performed using a subject-specific trained long short-term memory network (LSTM).

In order to overcome the issue of imbalanced data due to the lower frequency of occurrence of *stim* trials, especially the challenges of learning from the less
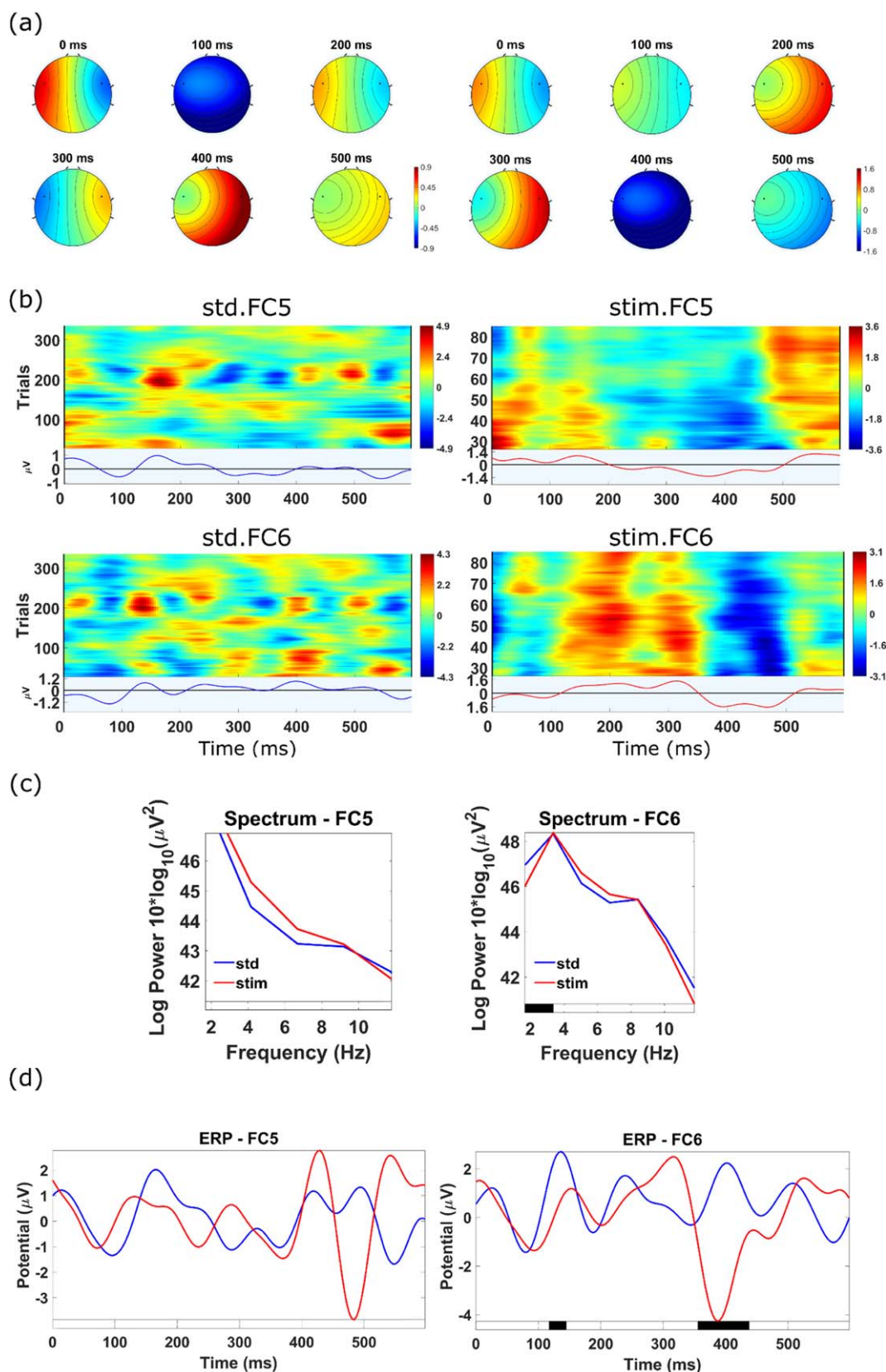
Figure 3. Error-related potentials (ErrPs) in a choice reaction task designed to elicit a feedback-related negativity (RFN). (a) 2D ErrP scalp maps at electrodes FC5 and FC6 for correct (*std*) and error (*stim*) trials. (b) Single-trial potentials in both trial conditions were plotted as a colour-encoded image (ErrP images) with a smoothing window of 50 trials, as in [24], in the correct (*std*, n = 359 epochs from 3 subjects) and error (*stim*, n = 109 epochs from 3 subjects) trials. $t = 0$ corresponds to the feedback presentation onset. (c) Power spectrum comparison between correct (*std*) and error (*stim*) trials at electrodes FC5 and FC6. The results of a sample-by-sample 1-way ANOVA with false discovery rate correction for multiple comparisons is depicted below the plot, with black segments depicting statistically significant ($p < 0.05$) differences between the power spectrum curves. (d) Grand average ErrPs at electrodes FC5 and FC6 for correct (*std*) and error (*stim*) trials time-locked to the onset of feedback presentation. $t = 0$ corresponds to the feedback presentation onset. The results of a sample-by-sample 1-way ANOVA with false discovery rate correction for multiple comparisons is depicted below the plot, with black segments depicting statistically significant ($p < 0.05$) differences between the grand average curves.

represented class, we provided only balanced data to the LSTM for training, by artificially multiplying *stim* trials to match the number of *std* trials.

For training, the dataset was split randomly into 80% epochs for training *per se* and 20% epochs for training validation. These epochs should not be confused with the signal epochs, because in the context of the training of an artificial neural network one epoch designates the running of an entire dataset through the neural network once. In this sense, the LSTM network was in as much efficient in that a training run of 50 epochs took around 8.5 min per subject to complete using minimalistic resources (a customer-grade laptop CPU Intel® i5 with 4 cores @1.60 GHz).

On unexpanded datasets (*orig*), the LSTM managed an average classification accuracy of (mean ± s.e. m.) 72.64% ± 1.34% (FC5) and 73.43% ± 1.34% (FC6), whereas classical methods such as LDA and QDA achieved lower, but statistically similar classification accuracies: FC5 LDA 64.21% ± 4.29% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 1.11, P = 0.14$), FC6 61.86% ± 4.09% LDA (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 1.53, P = 0.06$) and FC5 QDA 66.84% ± 3.84% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.91$, $P = 0.18$), FC6 QDA 66.43% ± 4.05% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.89, P = 0.19$) (figure 4(b), table 1).

However, an augmentation of the training datasets of factor 1 using DGCANs (*orig50dcgan50*) significantly improved classification accuracies compared to non-augmented datasets (*orig*) classified with the LSTM: FC5 *orig50dcgan50* 76.5% ± 1.33% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.00$, $P = 0.02$), and FC6 77.93% ± 1.17% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.89$, $P = 0.03$) (figure 4(b), table 1).

This was also the case with a factor 1 augmentation using IMFs (*orig50imf50*) compared to non-augmented datasets (*orig*): FC5 *orig50imf50* 77.29% ± 1.14% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.19$, $P = 0.01$) and FC6 *orig50imf50* 80.86% ± 1.86% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.58, P = 0.01$) (figure 4(b), table 1).

Expansion of the training dataset of the order 1 using either of DGCANs (*orig50dcgan50*) or IMFs (*orig50imf50*) yielded no difference in classification performance: FC5 Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.19$, $P = 0.42$ and FC6 Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.68, P = 0$. (figure 4(b), table 1).

Experimenting with a mix of the two augmentation methods did not lead to improved results compared to a single augmentation paradigm. Firstly, we tried a serial augmentation by initially expanding the training dataset by a factor of 1 using IMFs and then expanding the augmented dataset by another factor of 1 using DCGANs (*orig25imf25dcgan50*). In this way, the DCGANs also used IMF-generated epochs for learning. However, accuracies did not improve significantly compared to the non-augmented datasets: FC5 *orig25imf25dcgan50* 74.57% ± 1.27% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.03$, $P = 0.15$) and FC6 *orig25imf25dcgan50* 75.00% ± 1.29% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.41, P = 0.34$). Secondly, a parallel augmentation by a factor 1 using both methods simultaneously (*orig33imf33dcgan33*) also could not improve classification accuracy significantly when compared with the non-augmentation paradigm: FC5 *orig33imf33dcgan33* 71.14% ± 1.5% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.34$, $P = 0.37$), FC6 *orig33imf33dcgan33* 72.00% ± 2.15% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.52$, $P = 0.30$) (figure 4(b), table 1).

Next, we inquired as to why the classification accuracy was enhanced in the augmentation paradigms *orig50imf50* and *orig50dcgan50* compared to classical classification methods and then to the non-augmented paradigm.

Interestingly, when compared to classical classification methods, although classification sensitivity was marginally, but not significantly higher in the *qda* paradigm—FC5 *qda* 46.71% ± 9.02% *versus orig50imf50* 41.86% ± 5.08% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.2$, $P = 0.42$), *qda* versus *orig50dcgan50* 42.86% ± 6.92% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.004$, $P = 0.50$); FC6 *qda* 54.50% ± 6.87% versus *orig50imf50* 47.07% ± 6.51% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.82$, $P = 0.21$), *qda* versus *orig50dcgan50* 50.64% ± 3.96% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.19, P = 0.42$) (table 2) — the classification specificity was significantly higher in the *orig50imf50* and *orig50dcgan50* paradigms — FC5 *orig50imf50* 89.64% ± 2% versus *qda* 71.93% ± 4.61% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 2.84$, $P = 0.002$), *orig50dcgan50* 87.21% ± 3.16% versus *qda* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 2.38$, $P = 0.01$); FC6 *orig50imf50* 90.86% ± 2.11% versus *qda* 70.21% ± 5.21% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 3.19$, $P < 0.001$), *orig50dcgan50* 87.21% ± 1.81% versus *qda* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 1.89$, $P = 0.03$) (table 2). Another noteworthy aspect concerned improved classification precision in the LSTM-based paradigms compared to classical classification methods: FC5 *orig50imf50* 60.57% ± 5.76% versus *qda* 34.14% ±
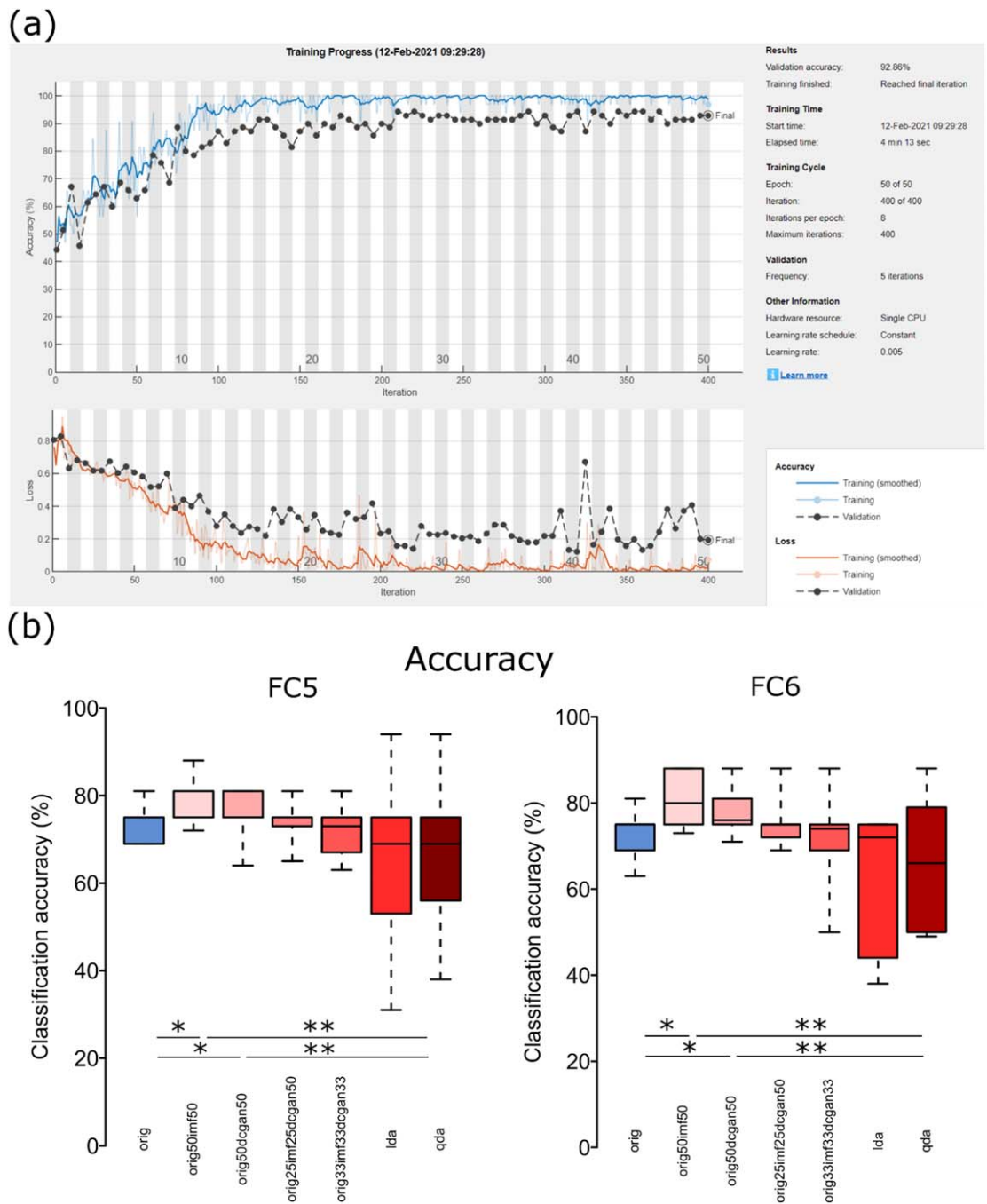
Figure 4. Online ErrP classification. (a) Example of a training plot of the LSTM network displaying the evolution of classification accuracy (above) and of the loss function (below) on the training (blue in plot above, orange in plot below) as well as on the validation dataset (black above as well as below). (b) Box plots of the mean $\pm$ s.e.m. classification accuracies depending on whether data augmentation is used and, if so, upon which method. *orig* = original, non-amplified training data; *orig50imf50* = 1:1 training data augmentation using IMF, so that 50% of the training set represent original, 50% represent artificial data; *orig50dcgan50* = 1:1 training data augmentation using DCGAN, so that 50% of the training set represent original, 50% represent artificial data; *orig25imf25dcgan50* = 1:1 training data augmentation using IMF, followed by a 1:1 amplification of the already amplified dataset using DCGAN; *orig33imf33dgcan33* = 1:2 training data augmentation using IMF as well as DCGAN in equal proportions; *lda* = linear discriminant analysis; *qda* = quadratic discriminant analysis. '*' statistically significant difference (Kruskal Wallis test with Dunn's correction) p < 0.05; '**' statistically highly significant difference (Kruskal Wallis test with Dunn's correction) p < 0.001.

5.1% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 3.11, P < 0.001$), *orig50dcgan50* 64.54% $\pm$ 7.19% versus *qda* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 3.12, P < 0.001$), FC6 *orig50imf50* 67.71% $\pm$ 6.13% versus *qda* 43.71% $\pm$ 6.03% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 2.77,$

$P = 0.003$), *orig50dcgan50* 58.79% $\pm$ 3.79% versus *qda* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 2.28, P = 0.01$) (table 2).

  When compared to the non-augmented paradigm, the results were similar to the case of classical classification algorithms, with marginally higher precision — FC5 *orig50imf50* 60.57% $\pm$ 5.76% versus

Table 1. Mean $\pm$ s.e.m. online classification accuracy.

Dataset type	Electrode	LSTM	LDA	QDA
orig	FC5	72.64% $\pm$ 1.34%	64.21% $\pm$ 4.29%	66.84% $\pm$ 3.84%
	FC6	73.43% $\pm$ 1.34%	61.86% $\pm$ 4.09%	66.43% $\pm$ 4.05%
orig50dcgan50	FC5	76.5% $\pm$ 1.33%		
	FC6	77.93% $\pm$ 1.17%		
orig50imf50	FC5	77.29% $\pm$ 1.14%		
	FC6	80.86% $\pm$ 1.86%		
orig25imf25dcgan50	FC5	74.57% $\pm$ 1.27%		
	FC6	75.00% $\pm$ 1.29%		
orig33imf33dcgan33	FC5	71.14% $\pm$ 1.5%		
	FC6	72.00% $\pm$ 2.15%		

Table 2. Mean $\pm$ s.e.m. online classification sensitivity, specificity and precision.

Parameter	Electrode	LSTM *orig50dcgan50*	LSTM *orig50imf50*	LSTM *orig*	QDA
Sensitivity	FC5	42.86% $\pm$ 6.92%	41.86% $\pm$ 5.08%	42.21% $\pm$ 4.75%	46.71% $\pm$ 9.02%
	FC6	50.64% $\pm$ 3.96%	47.07% $\pm$ 6.51%	46.21% $\pm$ 5.16%	54.50% $\pm$ 6.87%
Specificity	FC5	87.21% $\pm$ 3.16%	89.64% $\pm$ 2%	82.64% $\pm$ 2.44%	71.93% $\pm$ 4.61%
	FC6	87.21% $\pm$ 1.81%	90.86% $\pm$ 2.11%	80.64% $\pm$ 1.79%	70.21% $\pm$ 5.21%
Precision	FC5	64.54% $\pm$ 7.19%	60.57% $\pm$ 5.76%	50.79% $\pm$ 5.35%	34.14% $\pm$ 5.1%
	FC6	58.79% $\pm$ 3.79%	67.71% $\pm$ 6.13%	43.71% $\pm$ 2.73%	43.71% $\pm$ 6.03%

orig 50.79% $\pm$ 5.35% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.29$, $P = 0.1$), *orig50dcgan50* 64.54% $\pm$ 7.19% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.33$, $P = 0.1$); FC6 *orig50imf50* 67.71% $\pm$ 6.13% versus *orig* 43.71% $\pm$ 2.73% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.82$, $P = 0.002$), *orig50dcgan50* 58.79% $\pm$ 3.89% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.32$, $P = 0.01$) — and specificity — FC5 *orig50imf50* 89.64% $\pm$ 2.00% versus *orig* 82.64% $\pm$ 2.44% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.63$, $P = 0.05$), *orig50dcgan50* 87.21% $\pm$ 3.16% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.18$, $P = 0.12$); FC6 *orig50imf50* 90.86% $\pm$ 2.11% versus *orig* 80.64% $\pm$ 1.79% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -2.93$, $P = 0.002$), *orig50dcgan50* 87.21% $\pm$ 1.81% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -1.63$, $P = 0.05$) — in the augmented paradigms, while sensitivity was only marginally, but not significantly improved: FC5 *orig50imf50* 41.86% $\pm$ 5.08% versus *orig* 42.21% $\pm$ 4.75% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.24$, $P = 0.41$), *orig50dcgan50* 42.86% $\pm$ 6.92% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = 0.48$, $P = 0.48$); FC6 *orig50imf50* 47.07% $\pm$6.51% versus *orig* 46.21% $\pm$ 5.16% (Kruskal-Wallis test followed by Dunn's multiple comparisons test, $H = -0.33$, $P = 0.49$), *orig50dcgan50* 50.64% $\pm$ 3.96% versus *orig* (Kruskal-Wallis test followed by Dunn's multiple

comparisons test, $H = -0.66$, $P = 0.25$) (table 2). Between the two alternative augmentation methods, IMF and DCGAN, there was no significant difference regarding subsequent classification accuracy, precision, sensitivity or specificity.

4. Discussion

In this study, we present a viable deep-learning augmented ErrP discriminating BCI using a consumer-grade portable headset EEG (Emotiv EPOC$^+$) as a proof-of-concept for more portable, artificial intelligence-enhanced and efficient BCIs that can transgress the barriers of the controlled scientific laboratory and offer a practical solution for patients with permanent neurological disabilities.

We find this approach interesting at least for several reasons. Firstly, demonstrating reliable ERP and, in particular, ErrP detection using a wireless headset and minimalistic computational resources mitigates the necessity of using professional EEG recording systems that constrain the experimental paradigm to the niche of a scientific laboratory. Secondly, an ErrP-based interface offers a convenient binary coverage of the logical space in a choice task, allowing to theoretically decipher a complex intention by splitting it in multiple successive binary subtasks. Thirdly, deep-learning-based dataset augmentation can enhance classification performance and shorten training time, thus increasing the applicability range of the BCI and reducing the burden on potential patients.

In spite of the relative simplicity of the Emotiv EPOC$^+$ recording setup, we were able to demonstrate reliable offline ErrP detection. This was inferred by the

shape and timing of the grand averages exhibiting the canonical FRN during *stim* epochs. Consistent with the choice of electrodes, we noticed an up to 100 ms higher latency of the FRN than the literature reports, since most data has been gathered with 32-/64-electrode professional EEG systems that permitted the placement of electrode Cz. The recorded latency in our case could be due to the time necessary for propagation of the ErrP-induced dipole on the cortex surface up to the recording sites FC5 and FC6. We also observed consistent results on both FC5 and FC6 electrodes in spite of inherently more frequent frontal muscle artefacts than on Cz.

For online signal classification we trained an LSTM consisting of a bidirectional LST layer fed into a fully connected layer, followed by a *softmax* layer and finally a classification layer. We achieved significantly higher classification accuracies (in average 73%) than was the case with classical machine learning methods such as LDA (in average 64%) and QDA (in average 66.5%) and which were slightly inferior or comparable to previously published models based on ErrPs [40].

In analogy to the idea in modern computer vision deep learning paradigms, we hypothesized that data augmentation via either DCGANs or IMFs could increase accuracy by providing a broader training spectrum without the necessity for longer recording times. Augmenting the training dataset by a factor of 1 (50% original epochs, 50% artificial epochs) proved to be the most optimal compromise between maintaining enough original data and providing sufficient variability through augmented epochs. In this way, we achieved classification accuracies between 77% and 81%, with no significant difference between augmentation using DCGANs or IMFs. Further increase of the augmentation factor did not lead to significant improvement but led to an increase in LSTM training time. Improvement in classification accuracy upon dataset augmentation was paralleled by increased precision and specificity, but not increased sensitivity in the detection of *stim* epochs.

Future work could focus on how similar models behave in first virtual reality and then more elaborate practical settings, exploiting the reliable binarity offered by the decoding of ErrPs. One drawback of our study was the lack of a specialized Cz electrode, due to the constraints imposed by a prefabricated EEG recording device not primarily designed for recording ErrPs. It would be of note that previous studies have shown reliable classification of ErrPs in the context of choice reaction tasks. For instance, one study reached high classification accuracies between 86% and 96% in a sample of 9 subjects using an online adaptation method for an SVM classifier [41]. In another study across 12 subjects involved in binary choice tasks, accuracies between 63% and 76% were achieved depending on the exact task [42]. Finally, in a human-robot interaction study, ErrPs were decoded online

with an average accuracy of 82% across 13 subjects [40]. However, all these studies have used professional EEG recording equipment with at least 32 electrodes, did not employ wireless headsets and also did not involve dataset augmentation methods.

5. Conclusion

In this study, we described a deep-learning enforced ErrP discriminating BCI using novel data augmentation methods analogous to those employed in computer-aided vision algorithms using consumer-grade hardware resources including a portable headset EEG. We achieved online discrimination accuracies of between 76% and 81% across 14 subjects which were comparable to those obtained using 32/64-channel professional EEG devices by coupling DCGAN- or IMF-based dataset augmentation to online classification via an LSTM neural network. Our study is, to our knowledge, the first to combine a wireless EEG headset with deep-learning algorithms for data expansion and classification to generate a portable and practical EEG-based BCI that has the potential to expand the applicability of BCIs beyond the niche of a scientific laboratory.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

ORCID iDs

Mihai Ancau ⓘ https://orcid.org/0000-0002-7454-7178

References

[1] Kandel E R, Schwartz J H, Jessell T M, Siegelbaum S A and Hudspeth A J 2013 *Principles of Neural Science* ed E R Kandel 5th edn (New York: McGraw Hill)

[2] Thielen J, Marsman P, Farquhar J and Desain P 2021 From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface *J. Neural. Eng.* (https://doi.org/10.1088/1741-2552/abecef)

[3] van Dinteren R, Arns M, Jongsma M L and Kessels R P 2014 P300 development across the lifespan: a systematic review and meta-analysis *PloS. One* **9** e87347

[4] Rabie A 2017 Ramadan AVV. Brain computer interface: control signals review *Neurocomputing* **223** 26–44

[5] Lievesley R, Wozencroft M and Ewins D 2011 The Emotiv EPOC neuroheadset: an inexpensive method of controlling assistive technologies using facial expressions and thoughts? *Journal of Assistive Technologies* **5** 67–82

[6] Debener S, Minow F, Emkes R, Gandras K and De Vos M 2012 How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology.* **49** 1617–21

[7] Duvinage M, Castermans T, Petieau M, Hoellinger T, Cheron G and Dutoit T 2013 Performance of the emotiv epoc headset for P300-based applications *Biomed. Eng. Online* **12** 1–15

[8] Ekanayake H 2010 P300 and Emotiv EPOC: Does Emotiv EPOC capture real EEG? Web publication. Accession date 2022-10-07.P300 and Emotiv EPOC: Does Emotiv EPOC capture real EEG? Web publication. Accession date 2022-10-07 (http://neurofeedback.visaduma.info/emotivresearch.htm)

[9] Ullsperger M, Danielmeier C and Jocham G 2014 Neurophysiology of performance monitoring and adaptive behavior *Physiol. Rev.* **94** 35–79

[10] Ferrez P W and JdR M 2008 Error-related EEG potentials generated during simulated brain–computer interaction *IEEE Trans. Biomed. Eng.* **55** 923–9

[11] Chavarriaga R, Sobolewski A and Millán J R 2014 Errare machinale est: the use of error-related potentials in brain-machine interfaces *Frontiers in Neuroscience.* **8** 208

[12] Zhang Q and Liu Y 2018 Improving brain computer interface performance by data augmentation with conditional deep convolutional generative adversarial networks *arXiv preprint arXiv* 180607108

[13] Radford A, Metz L and Chintala S 2015 Unsupervised representation learning with deep convolutional generative adversarial networks *arXiv preprint arXiv* 151106434

[14] Huang N E *et al* 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis *Proc. of the Royal Society of London Series A: mathematical, physical and engineering sciences* 454, 903–95

[15] Petruţiu V M *et al* (ed) 2020 Enhancing the classification of EEG signals using Wasserstein generative adversarial networks *2020 IEEE 16th Int. Conf. on Intelligent Computer Communication and Processing (ICCP); : IEEE* (https://doi.org/10.1109/ICCP51029.2020.9266157)

[16] Bellary S A S and Conrad J M (ed) 2019 Classification of error related potentials using convolutional neural networks *2019 9th Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence); IEEE* (https://doi.org/10.1109/CONFLUENCE.2019.8776901)

[17] Chavarriaga R, Iturrate I and Millán JdR 2016 Robust, accurate spelling based on error-related potentials *Proc. of the 6th Int. Brain-Computer Interface Meeting* (https://doi.org/10.3217/978-3-85125-467-9-15)

[18] Bollon J-M, Chavarriaga R, Millán J R and Bessiere P (ed) 2009 EEG error-related potentials detection with a Bayesian filter *4th Int. IEEE/EMBS Conf. on Neural Engineering* (IEEE) (https://doi.org/10.1109/NER.2009.5109393)

[19] Ventouras E M, Asvestas P, Karanasiou I and Matsopoulos G K 2011 Classification of error-related negativity (ERN) and positivity (Pe) potentials using kNN and support vector machines *Comput. Biol. Med.* **41** 98–109

[20] Ancău D, Roman N-M and Ancău M 2020 Analyzing error potentials using stacking algorithms *Journal of Medical and Biological Engineering.* **40** 575–81

[21] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *Science* **313** 504–7

[22] Alhagry S, Fahmy A A and El-Khoribi R A 2017 Emotion recognition based on EEG using LSTM recurrent neural network *Emotion.* **8** 355–8

[23] Yeung N, Holroyd C B and Cohen J D 2005 ERP correlates of feedback and reward processing in the presence and absence of response choice *Cerebral cortex.* **15** 535–44

[24] Iturrate I, Montesano L and Minguez J 2013 Task-dependent signal variations in EEG error-related potentials for brain–computer interfaces *J. Neural Eng.* **10** 026024

[25] Brainard D H and Vision S 1997 The psychophysics toolbox *Spatial Vis.* **10** 433–6

[26] Lin Y-P, Wang Y and Jung T-P 2014 Assessing the feasibility of online SSVEP decoding in human walking using a consumer EEG headset *Journal of neuroengineering and rehabilitation.* **11** 1–8

[27] Campbell A *et al* (ed) 2010 Neuro phone: brain-mobile phone interface using a wireless EEG headset *Proc. of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds* (https://doi.org/10.1145/1851322.1851326)

[28] Liu Y *et al* (ed) 2012 Implementation of SSVEP based BCI with Emotiv EPOC *2012 IEEE Int. Conf. on Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS) Proc.; IEEE* (https://doi.org/10.1109/VECIMS.2012.6273184)

[29] Güneysu A and Akin H L (ed) 2013 An SSVEP based BCI to control a humanoid robot by using portable EEG device *2013 35th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC); IEEE* (https://doi.org/10.1109/EMBC.2013.6611145)

[30] Bobrov P, Frolov A, Cantor C, Fedulova I, Bakhnyan M and Zhavoronkov A 2011 Brain-computer interface based on generation of visual images *PLoS One* **6** e20674

[31] Elliott C, Vijayakumar V, Zink W and Hansen R 2007 National instruments LabVIEW: a programming environment for laboratory automation and measurement *JALA: Journal of the Association for Laboratory Automation.* **12** 17–24

[32] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21

[33] Lopez-Calderon J and Luck S J 2014 ERPLAB: an open-source toolbox for the analysis of event-related potentials *Frontiers in Human Neuroscience* **8** 213

[34] Zhang K *et al* 2020 Data augmentation for motor imagery signal classification based on a hybrid neural network *Sensors* **20** 4485

[35] Lashgari E, Liang D and Maoz U 2020 Data augmentation for deep-learning-based electroencephalography *J. Neurosci. Methods* **346** 108885

[36] Goodfellow I *et al* 2014 Generative adversarial nets *Advances in neural information processing systems.* 10-Jun-2014 arXiv:1406.2661

[37] Zhang Z *et al* 2019 A novel deep learning approach with data augmentation to classify motor imagery signals *IEEE Access.* **7** 15945–54

[38] Alom M Z *et al* 2018 The history began from alexnet: A comprehensive survey on deep learning approaches *arXiv preprint arXiv* 180301164

[39] Clifford G D *et al* (ed) 2017 AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017 *2017 Computing in Cardiology (CinC)* (Piscataway, NJ: IEEE) (https://doi.org/10.22489/CinC.2017.065-469)

[40] Ehrlich S K and Cheng G 2018 Human-agent co-adaptation using error-related potentials *J. Neural Eng.* **15** 066014

[41] Spüler M, Rosenstiel W and Bogdan M 2012 Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning *PloS. One* **7** e51077

[42] Yousefi R, Sereshkeh A R and Chau T 2018 Exploiting error-related potentials in cognitive task based BCI *Biomed. Phys. Eng. Express* **5** 015023