Protocol design challenges in the detection of awareness in aware subjects using EEG signals

J. Henriques[1,2], D. Gabriel[3], L. Grigoryeva[1], E. Haffen[3,4,5] , T. Moulin[3,4,6,7], R. Aubry[3,8,9], L. Pazart[3], J.-P. Ortega[1,10,11]

Abstract

Recent studies have evidenced serious difficulties in detecting covert awareness with EEG based techniques both in unresponsive patients and in healthy control subjects. This work reproduces the protocol design in the two recent mental imagery studies (1,2) with a larger group comprising twenty healthy volunteers. The main goal is assessing if modifications in the signal extraction techniques, training-testing/cross-validation routines, and hypotheses evoked in the statistical analysis, can provide solutions to the serious difficulties documented in the literature. The lack of robustness in the results advises for further search of alternative protocols more suitable for machine learning classification and of better performing signal treatment techniques. Specific recommendations are made using the findings in this work.

Keywords: EEG, awareness detection, mental imagery, evoked potentials, EEG signal classification

1.- Introduction

Recent advances in brain imaging have led to the development of new methods for detecting awareness in unresponsive patients with disorders of consciousness (DOC). In the absence of obvious behavioral responses, imaging-based diagnostic methods can be used to reveal covert and volitional reactions. For example, some patients, initially diagnosed as being in a vegetative state, showed patterns of cerebral activation similar to those of healthy subjects when asked to perform mental imagery tasks like imagining playing tennis or moving inside their house, hence indicating the presence of awareness (3,4). Even though such findings may prove useful at the time of providing diagnoses and prognoses, they are still not part of standard clinical care. This is in part due to the fact that the detection methods developed so far require a good understanding of the implications of the protocols chosen and of the statistical techniques used to interpret the collected data since they may have a major impact on the results (5). Needless to say that inadequate analyses can lead to misdiagnoses and have major consequences for patients and their families.

[1] Laboratoire de Mathématiques de Besançon, Besançon, France.
[2] Cegos Deployment, Besançon, France.
[3] Centre d'investigation Clinique, CHU de Besançon, France.
[4] Laboratoire de Neurosciences, Besançon, France.
[5] Service de Psychiatrie de l'adulte, CHU de Besançon, France.
[6] Département de Recherche en imagerie fonctionnelle, CHU de Besançon, France.
[7] Service de neurologie, CHU de Besançon, France.
[8] Espace Ethique Bourgogne/Franche-Comté, CHU de Besançon/Dijon, France.
[9] Département douleur soins palliatifs, CHU de Besançon, France.
[10] Centre National de la Recherche Scientifique (CNRS).
[11] Corresponding author.

New electroencephalography (EEG) based protocols have been recently developed with the goal of assessing awareness (1,6–8). In (1) the authors demonstrated that it is feasible to use EEG signals acquired from patients with DOC by asking them to perform different mental tasks in order to determine the presence of awareness; in that specific study, the two mental tasks were imagining squeezing the right hand or imagining wiggling the toes. A part of the information extracted from the EEGs recorded while completing the two tasks was used for the supervised training of a support vector machine (SVM), whose performance was tested using the remaining data. The good accuracies in the testing phase that were detected for some patients in a vegetative or a minimally conscious state were declared in (1) as a sign of covert awareness. Unfortunately, a statistical reanalysis in (2) of the original data in (1) based on more realistic hypotheses, questioned the actual presence of covert awareness in these patients and, moreover, detected it in only 40% of the healthy control subjects, which is significantly lower than the 75% initially established. The authors of the original study answered in (9) that an analysis approach that is more sensitive but potentially prone to false detection is preferable under some circumstances to a poorly sensitive technique that prevents the appearance of false detection.

The magnitude of the discrepancies between these conflicting studies suggests the need to further investigate the robustness of the awareness detection results obtained with the chosen protocol. This is the main goal of our work, in which we tried to validate this detection method in aware subjects for which ground truth is known. More specifically, we analyzed the sensitivity of the protocol with respect to changes in the electrodes density, signal extraction technique used, choice of training-testing/cross-validation routines, and hypotheses evoked in the statistical tests that assess the significance of the obtained classification accuracies.

We assume in our work that only an awareness detection method that exhibits good performance when applied to fully aware subjects is likely to function properly in the detection of covert awareness. We hence replicated the EEG studies in (1,2) using a group comprising 20 healthy volunteers. The findings that we describe later on in the paper evidence a high variability in the obtained results but allow us to make specific recommendations for the design of improved protocols and better performing signal treatment techniques.

2.- Description of the study

2.1.- Protocol and participants

The participants were 20 neurologically healthy adults (11 female; 9 male) aged between 25 and 66 (mean 35.6). All the subjects were verified to be aware at the time of performing the task. The local ethical committee approved this study and all participants provided written informed consents.

The same command-following protocol as in (1) was replicated. Each volunteer had to perform mental imagery tasks of two types: (i) the right-hand task in which the subject has to imagine squeezing the right hand into a fist and then to relax it each time that a beep sounds, and (ii) the toes task, for which the instruction was to imagine wiggling the toes in both feet and then to relax them after a beep is heard. The pertinence of this imagery task in awareness detection is based on the results of numerous studies that show that, depending on the instruction, a sensorimotor cortical activity synchronization or desynchronization is

observed in the mu (8Hz-12Hz) and beta (13-30 Hz) bands (see (10–15) and references therein).

The experiment was carried out using a specific block structure. Each block contained 15 trials in during which the participants were completing the same imagery task. Each subject performed a total of 8 blocks of command-following tasks (4 blocks for the right-hand task and 4 blocks for the toes task) presented in a pseudorandomized order. A break was provided to participants before the onset of the next block. Each block was preceded by an auditory presentation of instructions explaining the task to be performed during the whole block each time that the participant would hear a beep. The 15 tones of each block were presented binaurally (600 Hz, 60 ms duration) with an inter-stimulus interval varying randomly between 4.5 and 9.5 seconds. The approximate average duration of the whole experiment for each participant was 120 minutes.

2.2.- Data acquisition and preprocessing

EEG signals were recorded using an OSG digital equipment (BrainRT; OSG bvba, Rumst, Belgium) with two Schwarzer AHNS epas 44 channels amplifiers (Natus, Munich, Germany). The signals were recorded with 64 electrodes at the positions of the 10/10 system using a 64 channel electrode cap (Easycap, EasycapGmbh, Ammersee, Germany); one of these electrodes was dedicated to the detection of ocular activity and was not directly used in the EEG signal acquisition. The sampling frequency was 1000 Hz. Data was high-pass filtered at 0.27Hz and no low-pass filter was applied. Signal preprocessing was carried out using the Cartool Software (http://brainmapping.unige.ch/Cartool.php). The EEG recordings consisted of 5500 ms long epochs that comprised 1500 ms of pre-stimulus and 4000 ms of post-stimulus signal that were extracted for each trial and participant. The baseline was defined using the 500 ms period preceeding the stimulus onset. Individual data were then recalculated against the average reference and band-pass-filtered to the frequency range 1-40 Hz. An automatic artifact detection procedure was then applied within the time period of interest that removed trials where an amplitude of more than 100 μV was present in at least one of the electrodes. Subsequently, the trials were visually inspected to remove eye blinks, movements, and muscular artifacts. Missing data due to artifacts were interpolated using a 3-dimensional spline algorithm (the average percentage of interpolated electrodes was 6.25%). The average number (±SE) of trials per subject accepted for the analysis was 110.4±6.3. The data used to conduct the study is available from the authors upon request.

2.3.- Signal extraction techniques under consideration

Three different signal extraction techniques, implemented via custom scripts in Matlab, were put to the test:

- **Fourier analysis based technique:** for comparison purposes, the method proposed in the original study (1) was reproduced. More specifically, among all the 63 cap electrodes only 18, that covered the motor area of the brain, were selected, namely: FC3, FC1, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6. Then, the acquired data signals were resampled at 100 Hz. The FFT (Fast Fourier Transform) coefficients were computed for the signals coming from each of the selected electrodes in the range between 7 and 30 Hz with a resolution of 0.39 Hz and using a sliding window of

1 s width with a resolution of 0.01 s. Since for each trial the tone was set at 1.5 s, the center of the first window was chosen at 2 s and for the last one at 5 s. A Hamming function was used before the estimation of each FFT. For each window we computed the average log-power values (natural logarithm of the absolute values of the FFT coefficients) over the four following frequency ranges: 7-13Hz, 13-19 Hz, 19-25Hz, and 25-30Hz. Finally, these average log-power values were normalized by using their mean and standard deviation across all windows, trials, and channels for a given subject; these normalized values were subsequently used in the classification procedure that we describe in detail in Section 2.4.

- **Extraction using parametric time series models:** this technique is customary in the construction of some brain computer interfaces (16–21). In our study, the EEG signals were extracted by fitting to them order three autoregressive models AR(3) (22). The resulting set of model coefficients (without the variance of the model driving noise) was subsequently used in the classification task. This approach was implemented using the signal coming from the 63 electrodes with a sampling frequency of 100 Hz. The whole sample comprising 5.5 s worth of signal was used for the estimation of the 63 sets of AR(3) coefficients.

- **Wavelets based technique:** this is another standard approach that is particularly well adapted to the non-stationary nature of the EEG signals. Examples of good results with this method in the EEG context can be found, for example, in (17,23–26) and references therein. In our experiment a wavelets based extraction was implemented by using the signals coming from the 63 electrodes and sampled at 100 Hz. For each channel, the wavelets expansion in terms of the Morlet family of the total sample length (5.5 s) were computed and the coefficients corresponding to a frequencies interval of 1-30 Hz with a frequency step of 0.25 Hz and a time step of 0.01 s were kept. The absolute values of these coefficients were then averaged on frequency ranges and sliding temporal windows identical to those used with the Fourier based technique. The frequency averaged coefficients for all the temporal windows and all the channels were used in the classification task. It is worth mentioning that other wavelet families were also considered but it is the Morlet one that offered the best performance.

2.4.- Classification: training-testing and cross-validation design

The classification of the trials for each of volunteers was carried out using a support vector machine (SVM) with a linear kernel. As in any machine learning procedure, the classification was implemented by splitting the available data into two sets. The first one, called training set, was used for the estimation of the SVM parameters (in other words, for training the SVM classifier) and the second one, called testing set, was put aside to evaluate the classification performance. The global accuracy of the classification was determined via a cross-validation strategy in which the performance was computed by averaging over different choices of partitions of the available trials into training and testing sets.

The data fed to the SVM was a concatenation of the different sets of parameters obtained for each trial after signal extraction (see Section 2.3), individually marked as belonging to one of the two classes of imagery tasks: squeezing the right hand into fist or wiggling the toes. All

the necessary procedures and data manipulations necessary for the classification task were implemented using custom scripts in Matlab.

The cross-validation design appears to be a delicate issue in this context. As it has already been pointed out in (2), the temporal dependence between the test-set blocks has an important influence on the evaluation of the classification accuracy. In the original work (1) the choice of the training and testing sets was done in the following way: two adjacent blocks of different types (one with right-hand imagery and one with toes imagery) were used for the testing data and the trials included in the remaining blocks served as the training set. The global classification accuracy was computed by averaging over all the available different combinations of distinct adjacent blocks in the testing phase. However, it was noticed in (8) that when the blocks used for testing are not contiguous anymore and the distance between them increases, the classification performance severely declines.

These findings showed the sensitivity of the evaluation accuracy with respect to the design of the cross-validation. In order to further study this phenomenon, we considered several training-testing block configurations and we also studied the impact of breaking the block structure by carrying out the cross-validation with randomized individual trials and not only blocks. More precisely, two types of cross-validation procedures were implemented:

(i) Block-based approach: it involves the randomization of entire blocks.
(ii) Trial-based strategy: the block structure is not preserved and single trials are randomly permuted.

For these two families of cross-validation routines we also studied the influence of respecting the chronology in the training-testing. More specifically, for the block-based approach three main configurations were considered (tests 1, 2, and 3 in the tables):

- Test 1: the testing set for the classification task consists of pairs of consecutive blocks of different types; the training set is made of the remaining blocks. This is the technique proposed in (1).
- Test 2: the testing set consists of pairs of blocks of different types, not necessarily consecutive; the training set is constructed out of the remaining blocks. This is analogous to the technique proposed in (2).
- Test 3: the training set is enlarged each time following the chronological order with two consecutive blocks of different types; the testing set is constructed using the two blocks that chronologically follow the blocks in the training set. We emphasize that in this case the chronological order is preserved both in the training and the testing data samples.

Regarding the trial-based strategy, we considered the following four configurations (tests 4, 5, and 6 in the tables):

- Test 4: the initial training set for the classification task consists of the first 20 trials; the testing set is the following trial (21st). The training set is then enlarged by one trial and the testing set is the following trial (22nd) and so on. The chronological order is preserved in the training-testing.
- Test 5: the training sets for the classification task are constructed with a sliding window containing 50 trials; the testing sets consist of single trials following each of the windows. The chronological order is preserved in the training-testing.

- Test 6: the trials are all randomly permuted and split into four groups. Each of these four groups is taken as testing set for the classification task while the remaining trials are used for the training. The average accuracy rate is obtained by repeating this operation 1000 times. The chronological order is hence not preserved in the training-testing.

2.5.- Statistical analysis and significance of the results

The ability of a given subject to adequately perform the requested imagery task is assessed by conducting a statistical test on the signal classification results whose H0 hypothesis is that the classification accuracy is 50%. The original study (1) invokes an independence hypothesis between trials (and hence between blocks) that allows the authors to declare that the number of correct answers follows under the null a binomial distribution with parameters 0.5 and the number of trials. As we discuss in detail later on and as it is already pointed out in (2), this independence hypothesis is empirically violated. We hence find it more appropriate to use a permutation test that consists of constructing an empirical distribution of accuracies evaluated for the different possible relabelings of the tasks associated to the blocks or the individual trials, depending on the type of cross-validation approach followed for each subject (see Section 2.4). With such a test, the probability of obtaining a better classification accuracy than the one attained when using the right labels serves as a sign of lack of statistical significance. Additionally, we apply a correction with respect to the false discovery rate (FDR) to the p-values coming from the permutation test (27). Even though for some of the cross-validation configurations the empirical distribution is constructed with a limited number of permutations, no unjustified independence hypotheses were invoked.

3.- Results

The results presented in Tables I and II show a considerable variability of the classification accuracy and of the statistical significance of the results with respect to modifications in the signal extraction technique and the cross-validation strategy. For example, the methods proposed in (1) (see the first column for Test 1 in Table I) produced, with the 20 subjects that participated in our study, a mean classification accuracy of 61.4% which allows us, using the FDR corrected permutation test to state that 65% (or 60% with the binomial test in that reference, not reported in the tables) of them were able to adequately perform the requested imagery task at the alpha=0.05 level. The corrections in the cross-validation design proposed in (2) make these figures (see the first column for Test 2 in Table I) dramatically drop to 55.8% and 15%, respectively. The differences that we just described regarding two specific methods are prevalent across all the tests carried out and techniques used.

---TABLES I AND II AROUND HERE---

The predominant results established in the study are:

Variability of the results with respect to the choice of cross-validation strategy. The results of the experiments indicate that the temporal dependence between the trials and between the blocks influences considerably the performance of the classification task. Indeed, the cross-validation design corresponding to the trial-based approach (tests 4, 5, and 6) substantially outperforms the block-based strategy (tests 1, 2, and 3). The best average accuracy rate obtained when respecting the blocks is 61.4% (Test 1 using the Fourier signal

extraction technique in (1)), while using a trial-based cross-validation we can reach classification performances up to 86.3% (Test 6 with AR(3) signal extraction). The same conclusion is drawn regarding the percentage of individuals whose classification performance indicates that they have successfully carried out the imagery task. According to the FDR corrected permutation test (alpha=0.05), the best result obtained when respecting the block structure is 65% (Test 1 with Fourier signal extraction) versus the 100% attained when using individual trials in several tests (Table II).

This phenomenon is due to the fact that individual trials are more statistically dependent within a block than across blocks. Indeed, as in the experiments 4, 5, and 6, the training and the testing sets are both likely to contain trials that are different but that belong to a same given block, the intra-block dependence would explain why the classification accuracy is higher.

Additionally, notice that the accuracy rates obtained in tests 4 and 5 are similar. The only difference between these two tests is that the training set in test 4 contains all the available past trials while in the case of test 5 it incorporates only the most recent 50 trials. The testing set for both instances includes one trial that chronologically follows the ones used for the training. The fact that including all the previous trials in test 4 does not help in improving the accuracy rate indicates that it is the dependence associated to the proximity between trials used in the training and the testing sets that is influential in obtaining a good performance.

--FIGURE 1 AROUND HERE--

In order to further explore this line of reasoning, we studied the changes of the classification accuracy when the training and testing sets contain or not trials that come from the same block. To be more specific, we start by taking as training set the first 35 trials for each subject and we then evaluate the classification accuracy using as testing set the trial number 36; next, we construct the training sample using the trials with numbers 2 to 36 and measure the accuracy rate with the trial number 37. We continue this procedure until we run out of trials. The graphs in Figure 1 represent the evolution of the classification error according to this construction for three different subjects and the three signal extraction techniques considered. The red bars in the graphs correspond to the first trial of each block and, in order to make the representation less volatile, each marker indicates the average classification error over the next 10 trials; for example, the marker at the position 36 indicates the average error obtained when classifying the trials from 36 to 45. In many instances, it is visible in these graphs that the classification error decreases right after the red bars. This phenomenon supports the intra-block dependence thesis because the markers in those positions have been computed using classification experiments that incorporate, simultaneously in the training and the testing sets, trials coming from a same given block.

Variability of the results with respect to the signal extraction technique, the EEG density, and its relation with the cross-validation. Our results show that the pertinence of a given signal extraction technique depends on the cross-validation scheme used. Indeed, the figures in the tables indicate that Fourier based methods are more appropriate than the AR(3) approach when working at the block level and, conversely, when we use individual trials in the cross-validation, the AR(3) models produce excellent performances that leave Fourier methods far behind. For example, even though the accuracies in the Test 1 obtained with the Fourier based technique and the AR(3) method are quite close (61.4% and 57.8%, respectively) the percentages of individuals whose performance of the requested task is

statistically significant according to the FDR corrected permutation test (alpha=0.05) are dramatically different, namely 65% and 10%, respectively. In contrast, these percentages go up to 85% and 100% in the trial-based Test 6 with associated average classification accuracies of 67.6% and 86.3%, respectively. Wavelets based extraction appears as a compromise between the Fourier and the AR(3) techniques.

It is worth mentioning that various other approaches to signal extraction were tested. We do not report the results obtained since they do not provide significant improvements. For example, in the context of the Fourier based techniques, we worked with the original sampling (1 KHz) of the signals with no resampling at 100 Hz, we used the absolute values of the spectrum instead of the log-power values in the classification, and we tried various high and low-pass frequency filters. As to the parametric time series approach, apart from the AR(3) model that we report on, we tested several other families: univariate ARMA (Autoregressive Moving Average (22)) and ARMA-GARCH (Autoregressive Moving Average – Generalized Autoregressive Conditional Heteroskedasticity (28)) processes as well as multivariate DCC (Dynamic Conditional Correlation (29)) type parametric models. Among all the tested specifications the autoregressive model AR(3) of order three demonstrated the best overall performance in the processing of the EEGs of the twenty available subjects.

--FIGURE 2 AROUND HERE--

The pertinence of a given EEG density seems also to be related to the cross-validation scheme used. Figure 2 shows the classification accuracy rates obtained for the different subjects when carrying out Test 1 (block-based cross validation) and Test 6 (trial-based cross validation) using the signal extraction techniques that prove to be the most successful for those experiments. The results show that the use of 63 electrodes systematically outperforms the 18 electrodes approach for trial-based cross validation schemes (this result is in agreement with the conclusions in (7)) and that the situation is right the opposite for block-based cross validation schemes. Unreported results show that this conclusion is independent of the signal extraction technique used.

Finally, we note that the preservation of the chronological order in the training-testing routines and in the associated cross-validation schemes has a weaker influence on the results than breaking or not the block structure.

4.- Discussion and recommendations

The goal of the replication of the protocol introduced in (1) and carried out in this paper was settling the debate about its validity in (2,9). The main criticism in (2) had to do with the correctness of the statistical analysis method used. The authors of the original study replied that their approach was certainly less conservative than the one proposed in (2) and argued about its pertinence in the evaluation of the presence of residual consciousness.

In that context, our results allow us to draw two main conclusions: 1) even when using the most advantageous statistical analysis methods, it is impossible to obtain perfect discrimination between tasks for all the subjects with the protocol proposed in (1). 2) This lack of performance prevails when using alternative signal extraction techniques and cross-validation strategies. These statements and the results that justify them allow us to formulate

suggestions on how to design future EEG mental imagery tasks to be used in unresponsive patients. We now discuss them.

A first point that needs to be understood is why no signal detection method leads to a perfect detection of awareness in all healthy participants whereas impressive results have been found with the same tasks in brain computer interface (BCI) protocols. An explanation for this apparent contradiction is that in BCI research, the discrimination between both imagery tasks is usually obtained after a long training period, which is not possible with unresponsive patients. Furthermore, as pointed out in (1), even after training the detection procedure does not still produce satisfactory results for a small proportion of subjects that are declared to be "BCI illiterates" (30). One can thus question the validity of using mental imagery with unresponsive patients. These patients may be able to understand but not to effectively perform the tasks for a variety of reasons.

Concerning the inconsistencies of the study design, we give strong evidence for the presence of intrablock dependence between trials, which poses a serious problem when using SVM as classification method. We recall that any machine learning based classification technique calls for an optimized similarity between the trials included in the training and testing sets. In the protocol that we implemented in this study, this resemblance takes place mainly among the trials inside a given block, probably due to the fact that the subjects know in advance that the same task has to be carried out for the entire duration of the block. This argument explains why the performance of the tests with a block-based cross-validation (experiments 1-3) is significantly lower than those that mix trials coming from different blocks (experiments 4-6). Recent mental imagery protocols that randomize the trials have shown performance improvements (7) but larger scale studies need to be implemented in order to confirm the pertinence of this approach.

In view of what we just discussed we propose three specific modifications in the experimental design that could improve the signal detection in unresponsive patients

1- Changing the tasks execution chronology. Our results evidence that the suppression of intrablock dependence is a prerequisite for the success of any future study on unresponsive patients. The intrablock dependence strongly alters the value of the data for machine learning classification, regardless the signal detection method used.

2- Changing the tasks themselves. Mental imagery protocols have shown good performances in the context of fMRI and BCI research; our results suggest that they might not be the most appropriate for EEG based evaluation of patients' consciousness. Other protocols used to measure awareness in unresponsive patients propose to focus rather on their attention abilities (8,31); promising results have been obtained with healthy volunteers and patients but a replication of such designs on a larger scale needs to be performed before considering them for use in clinical practice

3- Changing the emotional valence of stimuli to improve detection. An improvement of detection performances has been obtained when subjects had to imagine playing guitar rather than squeezing the right hand (32). Furthermore, performances were also improved when these subjects were familiar with this musical instrument. This research approach thus appears to be promising in order to evaluate patients' awareness, even though the specific familiarity of the unresponsive patient to be tested with the proposed tasks will be highly influential.

Conclusions and recommendations. The results of our study evidence excessive sensitivity of the mental awareness detection protocol introduced in (1) with respect to changes in the electrodes density, signal extraction technique used, training-testing/cross-validation routines, and hypotheses evoked in the statistical tests that assess the significance of the obtained classification accuracies. This feature has been observed when the protocol has been implemented with fully aware healthy volunteers.

The dramatic contrast in performances obtained with the various signal extraction methods and cross-validation designs shows very plausible intra-block dependence between trials that makes the experimental protocol inadequate for subsequent machine learning based classification.

In conclusion, our findings illustrate the difficulty of capturing awareness in fully aware healthy subjects with the EEG based protocols proposed so far and advise caution in their use at the time of detecting covert awareness in unresponsive patients. It is our opinion that as long as new classification techniques and clinical protocols that robustly detect awareness on healthy volunteers are not available, the use of EEG paradigms should be restricted to the evaluation of low-level cognitive processes at the patients' bedside. Only subjects showing signs of preserved cognitive functions would then benefit from a more expensive and complex fMRI evaluation.

Conflicts of interest

The authors declare that there are no conflicts of interest that could be perceived as prejudicing the impartiality of the research reported.

References

1. Cruse D, Chennu S, Chatelle C, Bekinschtein TA, Fernández-Espejo D, Pickard JD, et al. Bedside detection of awareness in the vegetative state: a cohort study. Lancet. 2011 Dec 17;378(9809):2088–94.

2. Goldfine AM, Bardin JC, Noirhomme Q, Fins J, Schiff ND, Victor JD. Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study."Lancet Corresp. 2013;381:289–91.

3. Monti MM, Vanhaudenhuyse A, Coleman MR, Boly M, Pickard, John D. Tshibanda L, Owen AM, et al. Willful modulation of brain activity in disorders of consciousness. N Engl J Med. 2010;362:579–89.

4. Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. Detecting awareness in the vegetative state. Science. 2006 Sep 8;313(5792):1402.

5. Phan TG. Disorders of consciousness: are we ready for a paradigm shift? Lancet Neurol. 2013 Feb;12(2):131–2.

6. Cruse D, Chennu S, Chatelle C, Fernández-Espejo D, Bekinschtein TA, Pickard JD, et al. Relationship between etiology and covert cognition in the minimally conscious state. Neurology. 2012 Mar 13;78(11):816–22.

7. Höller Y, Bergmann J, Thomschewski A, Kronbichler M, Höller P, Crone JS, et al. Comparison of EEG-features and classification methods for motor imagery in patients with disorders of consciousness. PLoS One. 2013;8(11).

8. Chennu S, Finoia P, Kamau E, Monti MM, Allanson J, Pickard JD, et al. Dissociable endogenous and exogenous attention in disorders of consciousness. NeuroImage Clin. 2013 Jan;3:450–61.

9. Cruse D, Gantner I, Soddu A, Owen AM. Lies, damned lies and diagnoses: estimating the clinical utility of assessments of covert awareness in the vegetative state. Brain Inj. 2014;28(9):1197–201.

10. Pfurtscheller G, Lopes da Silva F. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol. 1999;Nov. 110(11):1842–57.

11. Pfurtscheller G, Neuper C, Brunner C, Lopes da Silva F. Beta rebound after different types of motor imagery in man. Neurosci Lett. 2005;378(3):156–9.

12. Pfurtscheller G, Brunner C, Schlögl A, Lopes da Silva F. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. Neuroimage. 2006;31(1):153–9.

13. Pfurtscheller G, Scherer R, Müller-Putz G, Lopes da Silva F. Short-lived brain state after cued motor imagery in naive subjects. Eur J Neurosci. 2008;28(7):1419–26.

14. Graimann B, Huggins JE, Levine SP, Pfurtscheller G. Visualization of significant ERD/ERS patterns in multichannel EEG and ECoG data. Clin Neurophysiol. 2002;113(1):43–7.

15. Cruse D, Chennu S, Fernández-Espejo D, Payne WL, Young GB, Owen AM. Detecting awareness in the vegetative state: electroencephalographic evidence for attempted movements to command. PLoS One. 2012;7(11).

16. LaFleur K, Cassady K, Doud A, Shades K, Rogin E, He B. Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain-computer interface. J Neural Eng. 2013 Aug;10(4):046003.

17. Subasi A, Alkan A, Koklukaya E, Kemal Kiymik M. Wavelet neural network classification of EEG signals by using AR model with MLE preprocessing. Neural networks. 2005;18:985–97.

18. Anderson CW, Stolz EA, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng. 1998 Mar;45(3):277–86.

19. Rappelsberger P, Petsche H. Spectral analysis of the EEG by means of autoregression. Computerized EEG Analysis. 1975. p. 27–40.

20. Gersch W, Yonemoto J. Automatic classification of multivariate EEG's using an amount of information measure and the eigenvalues of parametric time series model features. Comput Biomed Res. 1977;10:113–25.

21. Jimenez JC, Biscay R, Montoto O. Modeling the electroencephalogram by means of spatial spline smoothing and temporal autoregression. Biol Cybern. 1995;72:249–59.

22. Brockwell PJ, Davis RA. Introduction to Time Series and Forecasting. Springer; 2002.

23. Petrosian A, Homan R, Prokhorov D, Wunsch II D. Classification of epileptic EEG using neural network and wavelet transform. Proceedings of SPIE The International Society for Optical Engineering. 1996. p. 834–43.

24. Adeli H, Zhou Z, Dadmehr N. Analysis of EEG records in an epileptic patient using wavelet transform. J Neurosci Methods. 2003 Feb 15;123(1):69–87.

25. Kiymik MK, Akin M, Subasi A. Automatic recognition of alertness level by using wavelet transform and artificial neural network. J Neurosci Methods. 2004 Oct 30;139(2):231–40.

26. Subasi A, Yilmaz M, Ozcalik HR. Classification of EMG signals using wavelet neural network. J Neurosci Methods. 2006 Sep 30;156(1-2):360–7.

27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.

28. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. J Econom. 1986;31(3):307–27.

29. Engle RF. Dynamic conditional correlation -a simple class of multivariate GARCH models. J Bus Econ Stat. 2002;20:339–50.

30. Guger C, Edlinger G, Harkam W, Niedermayer I, Pfurtscheller G. How many people are able to operate an EEG-based brain-computer interface (BCI)? IEEE Trans Neural Syst Rehabil Eng 2003; 11: 145-147. IEEE Trans Neural Syst Rehabil Eng. 2003;11:145–7.

31. Bekinschtein TA, Coleman MR, Niklison J, Pickard JD, Manes F. Can electromyography objectively detect voluntary movement in disorders of consciousness? J Neurol Neurosurg Psychiatry. 2009;79:826–38.

32. Gibson RM, Chennu S, Owen AM, Cruse D. Complexity and familiarity enhance single-trial detectability of imagined movements with electroencephalography. Clin Neurophysiol. Elsevier; 2013 Dec 13;

Caption for Figure 1: evolution of the classification error for three different subjects and the three signal extraction techniques considered as the training and testing sets are shifted by one trial according to the description in Section 4. The red bars in the graphs correspond to the first trial of each block. Each marker indicates the average classification error over the next 10 trials.

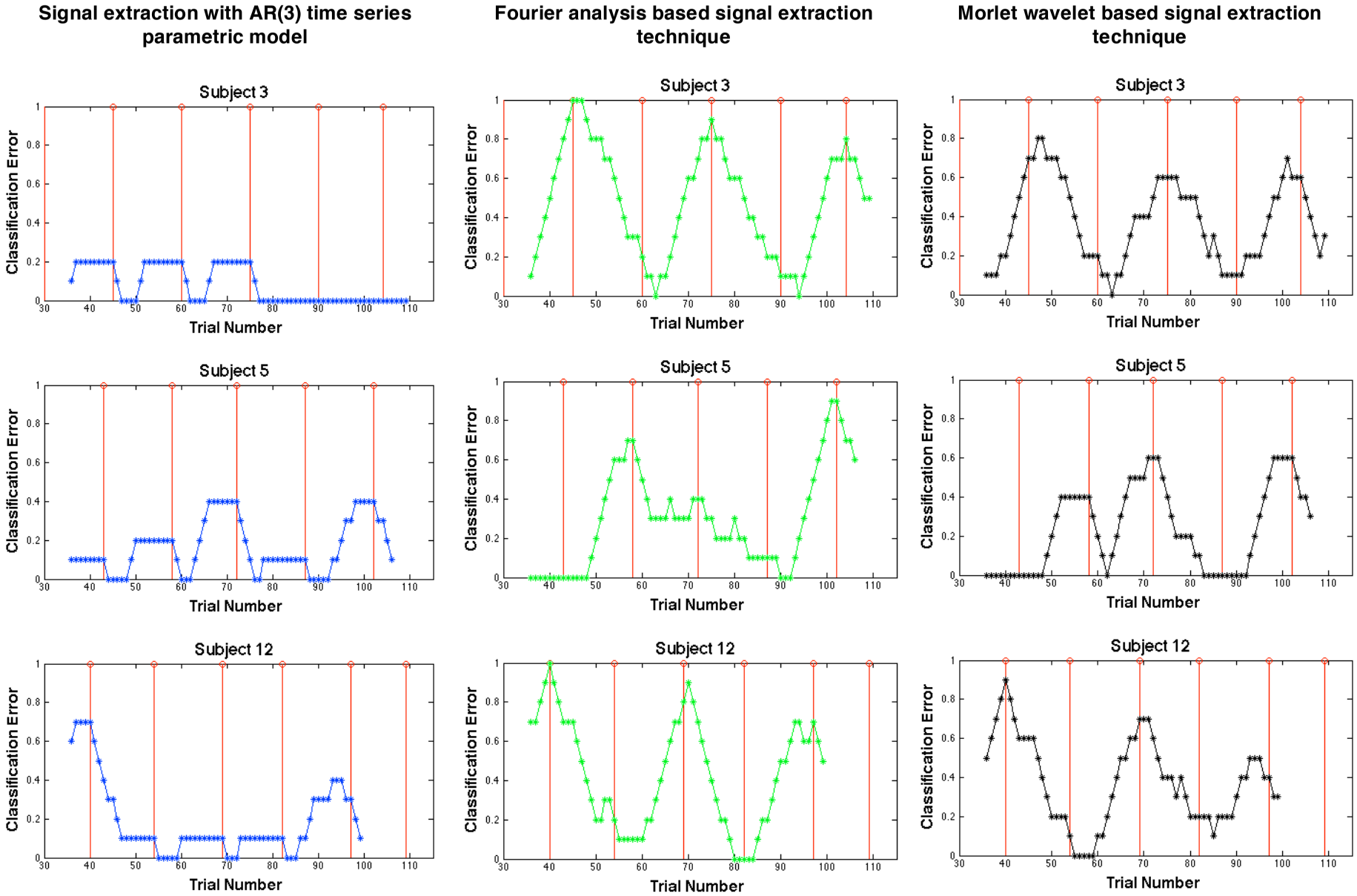
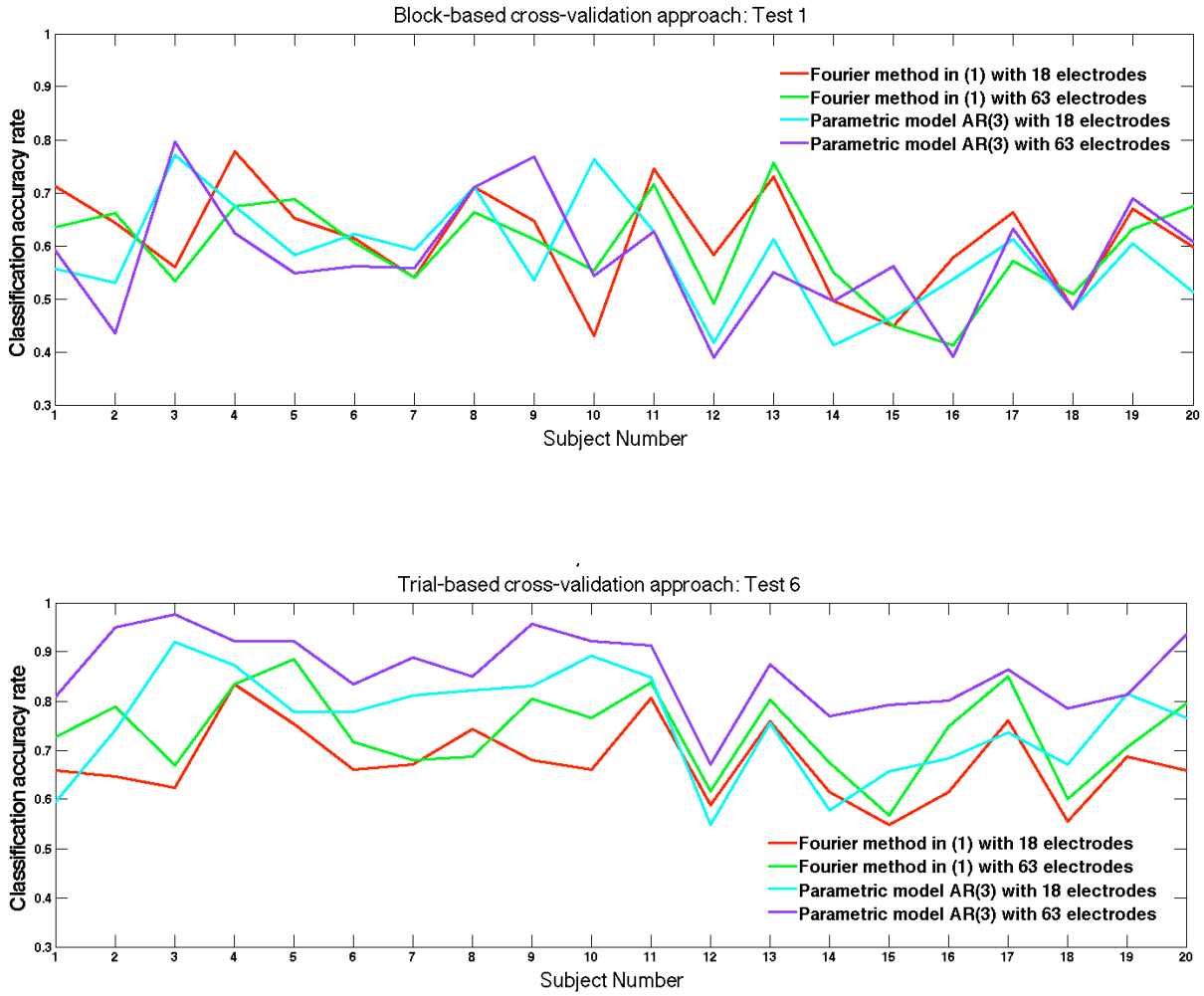

Signal extraction with AR(3) time series parametric model
Fourier analysis based signal extraction technique
Morlet wavelet based signal extraction technique
Subject 3
Subject 5
Subject 12
Classification Error
Trial Number

FIGURE 1

Caption for Figure 2: classification accuracy rates obtained for the different subjects when carrying out Test 1 (block-based cross validation) and Test 6 (trial-based cross validation) with 63 and 18 electrodes.

FIGURE 2



Caption for the tables: classification accuracy rates for the 20 healthy and aware subjects that took part in the study obtained by using each of the three EEG signal extraction techniques introduced in Section 2.3. Table I provides the results for tests 1, 2, and 3 in Section 2.4 that correspond to the block-based cross-validation design, and Table II reports on the tests 4, 5, and 6 that use trial-based cross-validation strategies. The row that follows the 20th subject provides the mean accuracy rate across the 20 subjects. The cells with colored background correspond to accuracy rates for which the p-values for the statistical test in Section 2.5 are smaller than the chosen significance level alpha=0.05 and for which we consider that the imagery task has been adequately performed. The row with caption "Success" reports the percentage of such accuracy rates for a given test and signal extraction technique.

Test number	1			2			3		
Cross-validation method	The testing set for the classification task consists of pairs of consecutive blocks of different types; the training set is made of the remaining blocks			The testing set for the classification task consists of pairs of blocks of different types, not necessarily consecutive; the training set is made out of the remaining blocks.			The training set for the classification task is enlarged each time with two consecutive blocks of different types; the testing set is constructed using the two blocks that follow the last pair of blocks in the training set. Chronological order is preserved in the training/testing.		
Subject No.	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate
1	0.713	0.591	0.583	0.598	0.424	0.511	0.693	0.545	0.580
2	0.643	0.435	0.678	0.598	0.476	0.613	0.576	0.482	0.624
3	0.559	0.797	0.644	0.489	0.748	0.621	0.551	0.708	0.697
4	0.778	0.624	0.726	0.718	0.645	0.684	0.761	0.716	0.659
5	0.652	0.548	0.678	0.574	0.357	0.535	0.563	0.563	0.701
6	0.614	0.561	0.649	0.623	0.526	0.605	0.671	0.471	0.694
7	0.540	0.558	0.460	0.407	0.431	0.316	0.442	0.570	0.500
8	0.710	0.710	0.645	0.654	0.430	0.458	0.675	0.638	0.588
9	0.647	0.767	0.586	0.603	0.606	0.539	0.640	0.616	0.581
10	0.430	0.544	0.640	0.414	0.480	0.559	0.356	0.437	0.448
11	0.745	0.627	0.784	0.760	0.561	0.711	0.701	0.506	0.714
12	0.583	0.389	0.435	0.528	0.271	0.363	0.537	0.439	0.488
13	0.730	0.550	0.757	0.676	0.538	0.743	0.753	0.556	0.790
14	0.495	0.495	0.596	0.438	0.388	0.454	0.543	0.469	0.519
15	0.448	0.562	0.400	0.305	0.393	0.260	0.333	0.474	0.321
16	0.577	0.392	0.443	0.541	0.389	0.469	0.554	0.432	0.432
17	0.663	0.633	0.602	0.592	0.495	0.602	0.542	0.431	0.500
18	0.482	0.482	0.418	0.382	0.355	0.298	0.500	0.488	0.438
19	0.670	0.689	0.585	0.637	0.675	0.500	0.588	0.638	0.475
20	0.598	0.607	0.709	0.618	0.528	0.658	0.551	0.584	0.652
Mean accuracy rate	0.614	0.578	0.601	0.558	0.486	0.525	0.576	0.538	0.570
Success	65%	10%	30%	15%	0%	5%	50%	15%	40%
Signal extraction technique	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets
No. of relabelings for permut. test	7			34			7		

TABLE I

Test number	4			5			6		
Cross-validation method	The initial training set for the classification task consists of the first 20 trials; the testing set is the following trial (21st). Training set is then enlarged by one trial and the testing set is the following trial (22nd) and so on. Chronological order is preserved in the training/testing.			The training sets for the classification task are constructed with a sliding window containing 50 trials; testing sets consists of single trials following each of the windows. The chronological order is preserved in the training-testing in this case.			The trials are all randomly permuted and split into four modules. Each of these four modules is taken as testing set for the classification task while the remaining trials are used for the training. The average accuracy rate is obtained by repeating this operation 1000 times.		
Subject No.	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate	Accuracy rate
1	0.653	0.768	0.716	0.692	0.738	0.677	0.659	0.810	0.768
2	0.568	0.895	0.842	0.585	0.938	0.938	0.647	0.950	0.858
3	0.622	0.959	0.714	0.588	0.926	0.750	0.624	0.976	0.770
4	0.773	0.887	0.753	0.776	0.940	0.716	0.833	0.923	0.852
5	0.747	0.821	0.789	0.523	0.846	0.631	0.754	0.923	0.883
6	0.660	0.798	0.649	0.641	0.781	0.594	0.661	0.835	0.731
7	0.677	0.860	0.710	0.635	0.841	0.730	0.672	0.889	0.744
8	0.678	0.839	0.690	0.684	0.912	0.754	0.744	0.850	0.789
9	0.625	0.948	0.813	0.788	0.970	0.864	0.680	0.958	0.856
10	0.660	0.894	0.777	0.813	0.984	0.859	0.660	0.921	0.843
11	0.744	0.841	0.866	0.673	0.788	0.865	0.807	0.914	0.900
12	0.602	0.761	0.670	0.690	0.931	0.724	0.589	0.671	0.603
13	0.747	0.846	0.846	0.738	0.770	0.852	0.758	0.875	0.825
14	0.584	0.708	0.697	0.695	0.814	0.729	0.615	0.769	0.713
15	0.529	0.788	0.600	0.473	0.818	0.618	0.549	0.792	0.632
16	0.610	0.766	0.675	0.617	0.787	0.660	0.614	0.800	0.745
17	0.756	0.795	0.782	0.708	0.750	0.792	0.761	0.865	0.882
18	0.556	0.711	0.578	0.583	0.750	0.600	0.555	0.786	0.640
19	0.698	0.802	0.640	0.661	0.679	0.571	0.687	0.814	0.724
20	0.598	0.918	0.784	0.627	0.896	0.866	0.658	0.937	0.901
Mean	0.654	0.830	0.729	0.676	0.843	0.740	0.676	0.863	0.783
Success	100%	100%	100%	75%	100%	95%	85%	100%	100%
Signal extraction technique	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets	Fourier method in reference (1)	Parametric model AR(3)	Morlet wavelets
No. of relabelings for permut. test	1000			1000			1000		

TABLE II