Comparing EEG Artifact Detection Methods for Real-World BCI

Michael W. Nonte[1(✉)], William D. Hairston[2],
and Stephen M. Gordon[1]

[1] Scientific Research Department, DCS Corporation,
Alexandria, VA, USA
{mnonte, sgordon}@dcscorp.com
[2] US Army Research Laboratory,
Human Research and Engineering Directorate,
Aberdeen Proving Ground, Aberdeen, MD, USA
william.d.hairston4.civ@mail.mil

Abstract. One major challenge to the real-world use of brain-computer interface (BCI) technology is the decrease in classifier performance caused by degradations in electroencephalogram (EEG) signal quality due to artifacts from non-neural electrophysiological activity and the gross movement of sensors and other EEG hardware. These artifacts can contaminate or mask the neural signal and thus cause a decrease in the performance of BCI classifiers due to the system's diminished ability to extract relevant features. One strategy to combat this effect is to identify and remove artifact-contaminated segments of data. We compared four methods that utilize higher order statistics to detect and artifact data on their ability to improve BCI classifier performance. We evaluated these methods on two datasets: a motor movement task and a rapid serial visual presentation (RSVP) task. In addition to comparing artifact detection methods, we compared the improvement in BCI classifier performance gained by removing artifact data to the decrease in performance caused by diminishing the amount of data available for classifier training. We found that overall the use of abnormal spectra to detect artifacts resulted in the greatest improvement to BCI classifier performance.

Keywords: Brain-Computer Interface (BCI) · Electroencephalography (EEG) · Artifact detection

1 Introduction

One crucial aspect of any brain-computer interface (BCI) system is a pre-processing pipeline that removes or mitigates non-neural signal components. This requirement is especially important when electroencephalography (EEG) is used to record neural activity, as the amplitude of the measured neural signal is small relative to electrophysiological and environmental noise. Generally, any recorded signal component from a non-neural source that is equal or larger in amplitude to the target brain-derived components is referred to as an artifact. These include non-neural electrophysiological

© Springer International Publishing Switzerland 2016
D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part I, LNAI 9743, pp. 91–101, 2016.
DOI: 10.1007/978-3-319-39955-3_9

activity as well as non-physiological sources of noise. Common physiological sources of artifacts include muscle activity, cardiac activity, eye blinks, and eye movement [1–3]. Non-physiological sources of artifacts include 50/60 Hz line noise, poor electrode contact, and cable sway [3]. Good BCI design attempts to control for and prevent artifacts, however, some artifacts such as eye blinks simply cannot be avoided, especially as research and BCI application moves out of confined settings and into more realistic, natural scenarios [4].

Artifacts can distort or mask the neurogenic signal in both the time and frequency domains [1]. BCI classifiers commonly extract time-amplitude or spectral features from EEG data to build a model distinguishing two or more different classes [5–9]. These classes may represent different behaviors, such as a left or right hand finger movement [5], or the neural responses to different stimuli, such as the presentation of a target or non-target image [6, 8]. The presence of artifacts can prevent a BCI classifier from building an accurate model, as the features extracted from contaminated training data may represent properties of the artifact rather than the underlying neural process. Additionally, the presence of artifacts in test data can cause misclassification even when the classification model is accurate. Thus it is beneficial to identify segments of data that are contaminated by artifact and remove them.

There exist multiple computationally efficient methods to identify segments of data contaminated by artifacts. These methods have typically been evaluated by computing a hit-rate using manually labeled artifact periods [2]. Using manually-labeled test data provides an evaluation criterion for a researcher looking to improve data quality for the sake of producing better statistical differentiation between experimental conditions or create a better visual representation of a neural process, but may not accurately inform the BCI developer about which artifact detection method will have the most positive impact on real-world BCI performance. For instance, it may be the case that the features used by a classifier are not affected by the presence of one or more types of artifacts, making the classifier resilient to their presence. Thus, the very definition of noise and artifact may be different to the BCI developer than it is for the researcher.

Additionally, the performance of a BCI classifier may actually decrease if the remaining data is insufficient to train an accurate model. That is, models can be overfit when data is too limited. There is a trade-off between improved performance due to the removal of artifact-contaminated data and a decrease in performance due to the reduction in the number of training samples; this tradeoff may differ depending on the BCI paradigm being used, the robustness of the classifier to noise, and the robustness of the classifier to a diminished training set.

In this work, we evaluate the ability of several popular artifact detection methods to improve the classification accuracy of common BCI classifiers. We compare these results to a case in which the training set size is held constant in order to distinguish the effects of removing artifact-contaminated data from that of reducing the size of the training set.

2 Background

2.1 BCI Classifiers

Three popular BCI classifiers were used for evaluation: common spatial patterns (CSP), hierarchical discriminant component analysis (HDCA), and xDAWN. CSP is designed for motor imagery and motor movement discrimination. CSP learns spatial filters to create components that maximize the ratio of variance between two task conditions, then uses the normalized log of variance of the component response as a feature for task discrimination [5, 9]. HDCA is a general-purpose classifier that is robust to temporal variability in the neural response. It divides data epochs into equal-sized, non-overlapping segments, trains a logistic regression classifier on each segment, then uses the output of each classifier to train a final logistic regression classifier that makes the final discrimination decision [6]. xDAWN was designed for oddball event detection and is commonly used for target detection in RSVP paradigms. It creates spatial filters to maximize the signal to signal plus noise ratio then trains a Bayesian linear discriminate analysis classifier on the resulting components [8].

2.2 Identifying Artifacts Using Signal Statistics and Spectral Power

One common method to deal with artifact-contaminated data is to simply remove it [1]. Prior to analysis, full datasets are commonly separated into equal-sized segments, called epochs, which are time-locked to events of interest. Delorme et al. [2] presented several methods using higher-order statistics, extreme values, and power spectral density to identify artifact-contaminated data. They showed that all of these methods were effective in identifying artifacts and that the detection of abnormal spectra was especially effective. We used four artifact detection methods presented in this work to rank the quality of data epochs: kurtosis, joint probability, extreme values, and abnormal spectra. Details of these methods can be found in [2, 10], but are paraphrased below.

Kurtosis. Kurtosis is a measure of the 'peakedness' of the probability distribution of a set of values. It is computed as the fourth standardized moment:

$$kurtosis(\boldsymbol{x}) = \frac{E\{(\boldsymbol{x} - \mu(\boldsymbol{x})^4\}}{(E\{\ (\boldsymbol{x} - \mu(\boldsymbol{x})^2\}\ )^2} \tag{1}$$

where $\boldsymbol{x}$ is the vector of data, *E{}* is the expectation operator and μ is the mean of the data. The kurtosis of the normal distribution is 3. Kurtosis values much lower than 3 are indicative of data that is mostly concentrated above and below the mean, with few values near the mean. This may reflect a process that varies rapidly between two values, such as an AC artifact, or a sudden change in signal amplitude offset, such as a mechanical movement of the electrode or an ocular artifact [2]. Kurtosis values much higher than 3 are indicative of data that is mostly concentrated close to the mean. This reflects a process in which the majority of the values are the same, such as in the case of a disconnected electrode. Thus, data with an excessively large or small kurtosis may contain an artifact.

Joint Probability. Another method to utilize the distribution of values within an epoch to detect artifacts is joint probability. In a broad sense, it computes the likelihood of observing the distribution of values in an epoch, given the distribution of values in the entire dataset. A probability density function (D_e) is computed for each electrode (e) using the entire dataset. Within each epoch (i), the joint log probability of values is computed for each electrode using:

$$J_e(i) = -\log\left(\prod_{x \in A_i} p_{D_e}(x)\right) \tag{2}$$

where $p_{De}(x)$ is the probability of observing the value x given D_e over all data in channel e and A_i are the values in epoch i.

Extreme Values. Neurogenic EEG signals are typically smaller than 100 µV in amplitude [1]. Large deviations in signal amplitude are then most likely the result of non-neural signal contamination. Thus, segments of data containing values much larger in amplitude than the rest of the dataset may contain an artifact.

Abnormal Spectra. Clean EEG has a frequency range of 0.01 to 100 Hz and has a power spectral density (PSD) that falls off roughly proportional to increasing frequency [1]. Some artifact types have characteristic spectral properties that cause abnormal deviations from the typical EEG PSD. For example, muscle artifacts have a large power concentration between 20–60 Hz and eye-related artifacts have a large power concentration between 1–3 Hz [2]. Segments of data displaying large increases in power amplitude in these frequency ranges relative to the rest of the data may contain an artifact.

3 Methods

We evaluated our chosen artifact detection methods on their ability to improve BCI performance using two different BCI paradigms as exemplar cases. The first data set was a finger movement study in which subjects performed self-paced movement of the middle and index fingers of both hands. The second data set was a rapid serial visual presentation (RSVP) study in which subjects were asked to detect targets of interest within a stream of rapidly presented visual stimuli.

3.1 Participants

We used data from 14 of 18 participants in a rapid serial visual presentation (RSVP) experiment and from 11 of 12 subjects in a motor movement experiment. Datasets containing excessive artifacts were deemed inappropriate for the current study and excluded. The investigators obtained the approval of the Institutional Review Boards of the Army Research Laboratory's Human Research and Engineering Directorate and adhered to Army policies for the protection of human subjects [11, 12].

3.2 Stimuli and Procedure

Insurgent-Civilian RSVP. Subjects were seated in front of a computer monitor and presented simulated images from a desert metropolitan environment. Images were presented at a rate of 2 Hz. In each image, if a person holding a gun was present it was considered a target image; if no humans were present in the image it was considered a non-target image. Subjects were instructed to attend to the presented images and count the number of target images. A total of 110 target images and 1346 non-target images were presented to each subject. For more information regarding this study, see [13].

Finger-Tapping. Subjects performed self-paced finger tapping movements using the middle or index finger of either hand. The time at which the downward movement of the finger was completed was recorded using a force-detecting switch. Subjects were instructed to leave between 4 and 5 s between successive taps. In each two minute block, the subject was told which finger to tap. The finger being tapped was changed on each trial. Subjects completed a total of 32 blocks so that each finger was used in 8 blocks.

3.3 Physiological Recording

Insurgent-Civilian RSVP. EEG data were recorded at 1024 Hz from 64 scalp electrodes using a BioSemi ActiveII system (Amsterdam, Netherlands). Channels were referenced offline using the average potential measured at two electrodes placed over the left and right mastoids. The data was bandpass filtered 0.1–50 Hz to reduce signal drift and high frequency noise.

Finger-Tapping. EEG data were recorded at 1024 Hz from 256 scalp electrodes using a BioSemi ActiveII system (Amsterdam, Netherlands). Channels were referenced offline using the average potential measured at two electrodes placed over the left and right mastoids. The data was bandpass filtered 0.1–50 Hz to reduce signal drift and high frequency noise.

3.4 Rejecting Epochs Based on High-Order Statistics

BCI Classification. CSP and HDCA were used to discriminate left from right hand finger movements in the finger-tapping dataset. Finger-tapping data was resampled to 128 Hz and RSVP data was resampled to 256 Hz. None of the scalp electrode channels were removed in either dataset prior to analysis. Continuous data was segmented into epochs around the event of interest. The event of interest for the finger-tapping data was defined as the detection of the downward movement of a finger based on the switches. Finger-tapping data was epoched -500 ms to 1500 ms relative to the event for HDCA and 500 ms to 1500 ms relative to the event for CSP; these epoch windows were determined to be optimal for each classifier based on a preliminary parameter search. The event of interest for the RVSP dataset was the onset of the presentation of a target or non-target stimulus image. RSVP data was epoched 0 to 500 ms relative to the event for all classifiers.

Epoch Ranking. Joint probability, kurtosis, extreme values, and abnormal spectra methods were used to rank epoch quality. The joint probability was computed using the EEGLAB function jointprob.m [14]. Built-in MATLAB (MathWorks, Natick, MA) functions were used to compute kurtosis and extreme values. These measures were computed for each channel within each epoch and normalized within channel across all epochs. Epochs were then ranked based on the absolute value of their normalized value, with larger values indicating a higher likelihood of containing an artifact.

The power spectral density of each channel within epoch was computed using the MATLAB function pmtm.m which implements a slepian multi-taper method to estimate the power spectral density (PSD) for the epoch. The mean PSD within each channel across all epochs was subtracted from each channel PSD estimate. For each epoch, the maximum spectral power (in dB) was found in the 0-2 Hz range and the 20–40 Hz range; these correspond to the frequency range of typical eye and EMG artifacts, respectively [1]. These two values were treated as a vector and the L^2 norm was used as a derived value to rank the epoch quality.

Cross-Validation. A 20-by-5 cross validation procedure was used to estimate classifier performance. Epochs are first randomly assigned to one of five partitions. Four of the partitions are used to train the classifier then the predictive accuracy of the model is tested on the remaining partition. The partition that is held out as the test set is rotated until all five partitions have served as the test set exactly once. The data is then repartitioned and the same procedure carried out again; this is repeated 20 times.

Epoch Rejection. For each subject, rejection methods, and dataset, a baseline performance value (in AUC) was computed using the entire dataset. Next, each method was evaluated by removing set percentages of data based on the ranking variables described above. In each case the data was removed then the cross-validation scheme was rerun to obtain a new estimate of the classifier performance. The percentage of data removed was incremented until the decrease in performance caused by reducing the training data available outweighed the increase in performance caused by the removal of artifact-contaminated data. In this case, as the percentage of data removed increases, the size of the training set naturally decreases by default; we will call this the dynamic training set (DTS) case.

To observe the effect of removing artifact-contaminated data in the absence of changes in training set size, we repeated this process using a fixed testing and training set size. Based on the results of the DTS case, we selected a percentage of data removal where the effect of the reduced training set size seems to equally counteract the effect of removing artifact data. We constrain the size of the training (and test) sets based on the size of the training and test sets for this level of data removal. For example, if we have 100 epochs, select 50 % as our removal constraint, and use an 80/20 split for training and testing data, our training and test set sizes will be set at 40 (100*0.5*0.8) and 10 (100*0.5*0.2), respectively. In the baseline (no data removal) case, we would then select 40 epochs for the training set and 10 epochs for the test set from the pool of 100 epochs on each iteration of the outer fold of the cross-validation. When we begin removing artifact data, we first remove it from the pool of available epochs, and again select 40 epochs for the training set and 10 epochs for the test set. This ensures that any change in classifier performance is attributable to the removal of artifact contaminated data rather than a change in the size of the training set. We will refer to this case as the fixed training set (FTS) case.

4 Results

To compare the effect of different percentages of data removal, we performed a 3-way ANOVA test using percentage of data removed, artifact detection method, and subject as the three grouping variables. We compared conditions using the increase in AUC relative to the mean baseline (no data removed) AUC. A multiple comparisons test was then performed using the Tukey-Kramer method with an alpha value of 0.05 to determine which percentages of data removal caused a significant increase or decrease in AUC over baseline. Note that error bars in Figs. 1, 2, 3 and 4 represent minimal group separation distances as computed by the multiple comparisons test and not standard deviation because this gives a better depiction using a more relevant statistic.

4.1 Finger Tapping

HDCA. Figure 1 shows the results of the multiple comparisons test for HDCA classification performance on the finger-tapping data. In the DTS case, the detrimental effect of reducing the training set size clearly outweighs the improvement gained by removal of artifact-contaminated data. The joint-probability, kurtosis, and abnormal spectra methods of artifact detection all cause a decrease in performance relative to

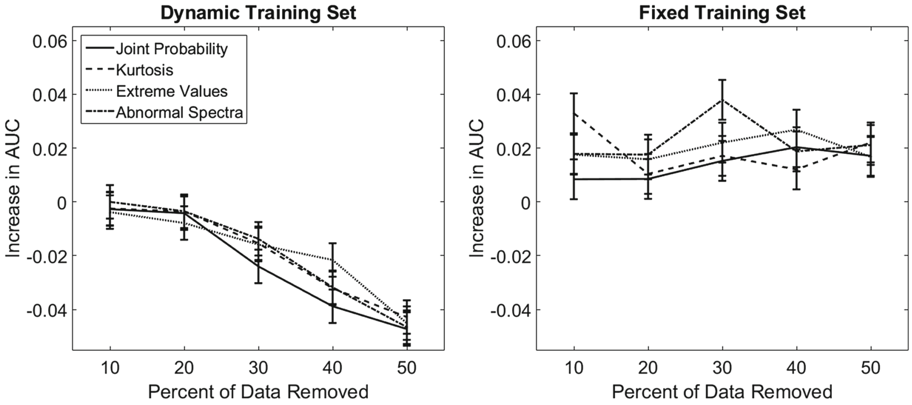

Fig. 1. Multiple comparisons for the effect of epoch rejection on HDCA performance in classifying right from left hand movements in a motor movement dataset. The effect of reducing the training set size significantly decreases AUC compared to baseline for all epoch rejection methods. In the dynamic training set (DTS) case, the size of the training and test sets decrease as the percentage of data removed increases. This decreases classifier performance and counteracts the positive effect of removing artifact data. In the fixed training set (FTS) case, the size of the training and test sets remain the same as the percentage of data removed increases. This control allows the effect of removing artifact data on classifier performance to be observed by itself.

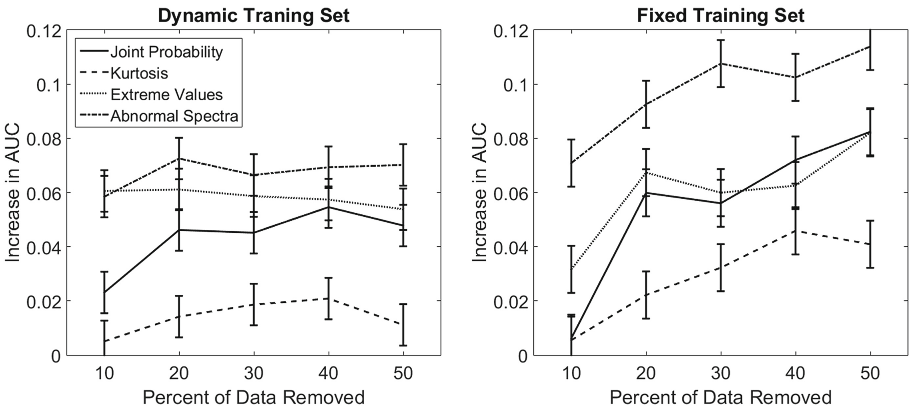

Fig. 2. Multiple comparisons for the effect of epoch rejection on CSP performance in classifying right from left hand movements in a motor movement dataset. The benefit of removing artifact contaminated data outweighs the detriment of decreasing the training set size.

baseline when 30 % or more of the data are rejected. The extreme values method of artifact detection performs even worse, showing a significant decrease in performance when 20 % or more of the data are rejected. In the FTS case, all rejection methods show an increase in classifier performance over baseline at all rejection percentage levels. Joint probability and extreme values show a weak upward trend in classifier performance, peaking at 40 % data removal, but the kurtosis and abnormal spectra methods do not show a relationship between data removal percentage and classifier performance.

CSP. Figure 2 shows the results of the multiple comparisons test for CSP classification performance on the finger-tapping data. In the DTS case, the improved classifier performance gained by the removal of artifact-contaminated data clearly outweighs the decrease in performance caused by decreasing the size of the training set. All methods cause improved classifier performance at all data rejection percentages, with the exception of a 10 % removal using the kurtosis method. In the FTS case, the joint probability and kurtosis artifact detection methods do not show a significant improvement in classifier performance when only 10 % of the data are rejected, but all other detection method and rejection percentage combinations show significant improvement in classifier performance over baseline.

4.2 Insurgent-Civilian RSVP

HDCA. Figure 3 shows the results of the multiple comparisons test for HDCA classification performance on the RSVP data. In the DTS case, the joint-probability and abnormal spectra artifact detection methods seem to overcome the decrease in performance due to decreased training set size. They show significant improvement over baseline at 25 % removal and above, and at 10 % removal and above, respectively. The kurtosis and extreme values methods do not cause a significant increase or decrease in

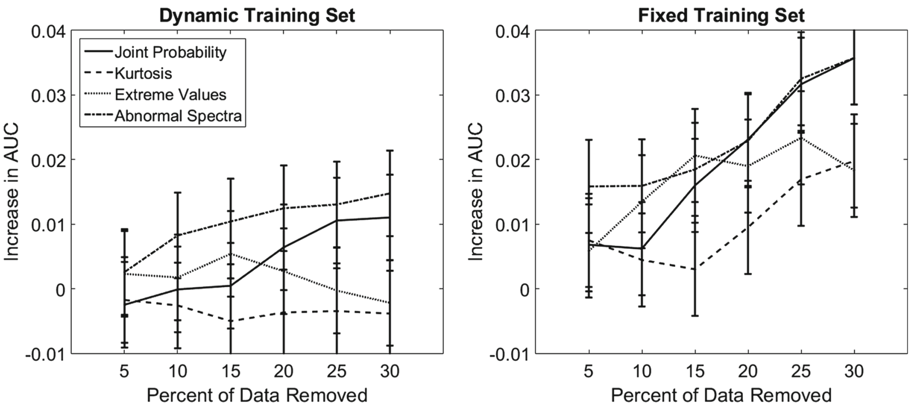

Fig. 3. Multiple comparisons for the effect of epoch rejection on HDCA performance in classifying target from non-target image presentations in an RSVP dataset.

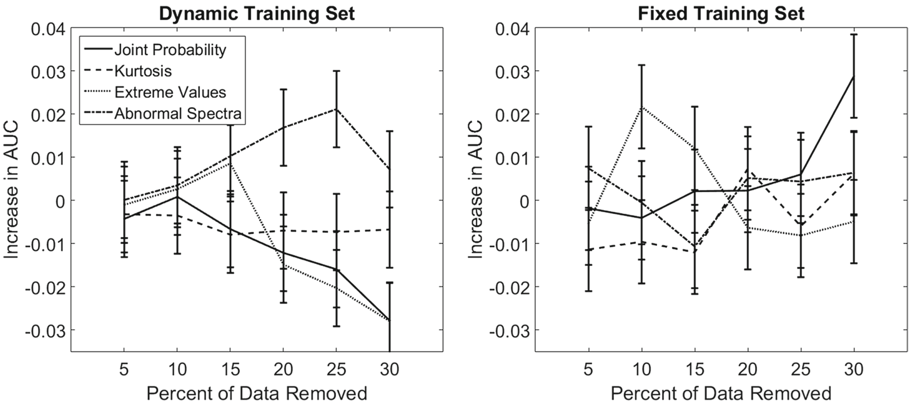

Fig. 4. Multiple comparisons for the effect of epoch rejection on xDAWN performance in classifying target from non-target image presentations in an RSVP dataset.

classifier performance. In the FTS case, all rejection methods show significant improvement over baseline. Performance improves over baseline for abnormal spectra at 5 % or greater data rejection, extreme values at 10 % or greater rejection, joint-probability at 15 % or greater rejection, and kurtosis at 20 % or greater rejection.

xDAWN. Figure 4 shows the results of the multiple comparisons test for xDAWN classification performance on the RSVP data. In the DTS case, the kurtosis method does not show a significant increase or decrease in performance. The joint-probability and extreme value methods show a significant decrease in classifier performance relative to baseline when 20 % or more of the data are removed. The abnormal spectra based rejection method shows a significant increase in performance from 20–25 % removal, but dips back below significance at the 30 % level. In the FTS case, the joint probability shows the only constantly increasing trend in classifier performance with increased data removal and only achieves a significant improvement at the 30 % level. Other rejection methods show a more irregular trend. Kurtosis-based rejection results in a significant decrease in performance at 5 % and 10 % rejection levels. Abnormal spectra based rejection results in a significant decrease at the 15 % rejection level. Finally, extreme value based rejection shows a significant increase in classifier performance at 10 % and 15 % rejection levels.

5 Discussion

Based on the results presented here, it is apparent that care must be taken to ensure a balance is met between the increase in classifier performance from the removal of artifact-contaminated data and the decrease in performance from reducing the training set size. In all cases other than RSVP classification with xDAWN, we see a clear difference in classifier performance between the fixed and dynamic training set cases, with the dynamic training set case consistently performing worse. This indicates that reducing the training set size does decrease classifier performance, but it can be seen that in many cases the benefit of removing artifact-contaminated data can overcome this detriment. CSP classification of the finger-tapping data and HDCA classification of the RSVP data seem to benefit in particular from epoch rejection with both showing a significant increase in performance when using most rejection methods. Interestingly, HDCA classification of the finger-tapping dataset shows a dramatic drop in performance when any of the detection methods are used to reject data. It is unclear at this point why HDCA classifier performance did not show this dramatic drop when applied to the RSVP data. Further study will be needed to determine if this difference in performance between motor and RSVP paradigms is consistent across other datasets.

In the cases where epoch rejection improved classifier performance, the abnormal spectra method consistently performed best. However, the current study only considered limited feature spaces, i.e. those captured by the classification methods HDCA, xDAWN, and CSP. When analyzing a new dataset, other methods of artifact detection should be considered, with the final selection tailored to the specific problem space. Based on our current results and the previous findings of Delorme et al. [2], we recommend the use of abnormal spectra as the starting point for improving BCI classifier performance.

In real-world BCI scenarios where the amount of training data is limited, careful consideration must be given as to how much data is rejected due to artifacts. The classifier-paradigm pairs studied here showed differences in their sensitivity to training set reduction and to artifact contamination. Further work must be done to develop an understanding of how to choose the best classifier for a paradigm, taking into consideration the amount of training data that will be available and the probability of artifact occurrence within that data.

Acknowledgements. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Urigüen, J.A., Garcia-Zapirain, B.: EEG artifact removal—state-of-the-art and guidelines. J. Neural Eng. **12**(3), 031001 (2015)
2. Delorme, A., Sejnowski, T., Makeig, S.: Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. Neuroimage **34**(4), 1443–1449 (2007)
3. Lawhern, V., Slayback, D., Wu, D., Lance, B.J.: Efficient labeling of EEG signal artifacts using active learning. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3217–3222. IEEE, October 2015
4. McDowell, K., Lin, C.T., Oie, K.S., Jung, T.P., Gordon, S., Whitaker, K.W., Hairston, W. D.: Real-world neuroimaging technologies. Access IEEE **1**, 131–149 (2013)
5. Müller-Gerking, J., Pfurtscheller, G., Flyvbjerg, H.: Designing optimal spatial filters for single-trial EEG classification in a movement task. Clin. Neurophysiol. **110**(5), 787–798 (1999)
6. Marathe, A.R., Ries, A.J., McDowell, K.: Sliding HDCA: single-trial EEG classification to overcome and quantify temporal variability. IEEE Trans. Neural Syst. Rehabil. Eng. **22**(2), 201–211 (2014)
7. Trejo, L.J., Rosipal, R., Matthews, B.: Brain-computer interfaces for 1-D and 2-D cursor control: designs using volitional control of the EEG spectrum or steady-state visual evoked potentials. IEEE Trans. Neural Syst. Rehabil. Eng. **14**(2), 225–229 (2006)
8. Rivet, B., Souloumiac, A., Attina, V., Gibert, G.: xDAWN algorithm to enhance evoked potentials: application to brain–computer interface. IEEE Trans. Biomed. Eng. **56**(8), 2035–2043 (2009)
9. Ramoser, H., Muller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehabil. Eng. **8**(4), 441–446 (2000)
10. Delorme, A., Makeig, S., Sejnowski, T.: Automatic artifact rejection for EEG data using high-order statistics and independent component analysis. In: Proceedings of the 3rd International Workshop on ICA, vol. 457, p. 462, December 2001
11. Regulation, A.: 70–25: Use of Volunteers as Subjects of Research. US Dept of the Army, Washington, DC (1990)
12. US Department of Health and Human Services. "Code of federal regulations: Protection of human subjects" (2011)
13. Cecotti, H., Marathe, A.R., Ries, A.J.: Optimization of single-trial detection of event-related potentials through artificial trials. IEEE Trans. Biomed. Eng. **62**(9), 2170–2176 (2015)
14. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods **134**(1), 9–21 (2004)