IOSR Journal of Mobile Computing & Application (IOSR-JMCA)
e- ISSN: 2394-0050, P-ISSN: 2394-0042.Volume 9, Issue 2 (Mar. – Apr. 2022), PP 06-13
www.iosrjournals.org

Discrimination of neuropsychiatric disease using EEG and Neurophysiological Biomarker Toolbox (NBT) with Machine Learning

Fayez Alshehri[1], Trent Lewis[2]
[1](*Ministry of Education, Riyadh, Saudi Arabia)*
[2](*College of Science and Engineering, Flinders University, Adelaide)*

Abstract:
Electromyogram (EMG) contamination has been shown to affect electroencephalogram (EEG) signals. Therefore, methods of isolating and removing EMG contamination are a focus of research. One of the most common ways to eliminate this contamination is through independent component analysis (ICA). Also, surface Laplacian (SL) has been proven to isolate the distant sources of EEG signals. The objective of this paper is to demonstrate the effects of EMG contamination on EEG signals using the Neurophysiological Biomarker Toolbox (NBT) and the impact of applying ICA, and ICA + SL on raw data. In this paper, the method for preparing the data is ICA with an auto-pruned method and SL using a flexible spherical spline. Machine learning was used to classify three neuropsychiatric diseases (anxiety, depression, and epilepsy) against control subjects under the three types of data pre-processing and raw data + SL. The data has been split into one second segments and classified according to features extracted from the NBT, which are the amplitude and the normalised amplitude for all frequency bands. Principal component analysis (PCA) was used for reducing the features, and 10-fold cross-validation and artificial neural networking were the methods that has been used for the classification. The results show a high percentage of accuracy in ICA + SL in all frequency bands. However, ICA in general has a percentage quite similar to the raw data, while SL, as well as ICA with a small percentage improved more than ICA and raw data. Overall, the gamma band for both amplitude and normalised amplitude in ICA + SL showed the best results, with accuracy over 87%, when comparing it with all disease classifications. Both results indicate that ICA + SL eliminate and isolate EMG contamination. However, the classification of ICA shows no significant change in the percentage of accuracy.
Key Word*: Electromyogram(EMG); electroencephalogram (EEG); Laplacian (SL); Machine learning; frequency bands.*

I. Introduction

Currently, the activities of the brain are non-invasively recorded with the help of an electroencephalogram, or EEG. An EEG offers exceptional temporal resolution and usability, which is why it is frequently used for brain-computer interface (BCI) research. BCI is a technology that offers differently abled people control over artificial communication and motor devices without the help of conventional mechanisms, such as nerves or peripheral muscles (Wolpaw et al. 2000; Bashashati et al. 2007).

It is important for a user to yield different patterns of brain activity to be able to control the EEG-based BCI. These patterns are recorded by electrodes that are attached to a person's scalp, and the outcomes are commands that are derived from algorithms and data that is mined from the EEG signals. As far as EEG signals are concerned, noise is ubiquitous because of functional variations and disparities present in the EEG, measurement inaccuracies, and elements like muscle movements and eye blinks. An unsuitable imaging of a motorised image-based BCI can also result in noise. The technologies for classification and extraction of features that are employed in BCIs are reviewed by Bashashati et al. (2007) and Garrett et al. (2003). Nonetheless, these elements can be eliminated, if ICA is used (Oja&Nordhausen 2001; Kachenoura et al. 2008), or excluded by criteria or thresholds.

On the other hand, ICA is a technique for processing signals that originated from blind source separation (Bell &Sejnowski 1995; Lee et al. 1999). Since then, ICA has frequently been applied in a number of fields, like speech processing, communication, and biomedical signal processing. ICA can decompose the observed multichannel signals into a number of autonomous constituents using an optimisation algorithm, which is driven by the principle of statistical independency. Neither of these techniques can identify the sound produced by incorrect selection of patterns of imaging because the information provided on the label is not considered (Sannelli et al. 2009). The ICA algorithm, on the other hand, needs visual inspection for the selection of artificial components that make its application impossible in an automatic BCI system.

Continuous EEG signals in clinical applications can be separated into numerous rhythms depending on their frequency: delta rhythm (0.3–4 Hz), theta rhythm (4–8 Hz), alpha rhythm (8–13 Hz), beta rhythm (13–30 Hz), and gamma rhythm (30–45 Hz). Cerebral diseases, such as cerebrovascular diseases, migraine and epilepsy, and EEG signals have a close correlation as the EEG of humans reflects the activity carried out by the nervous system. Hence, the method of processing and investigation of EEG signals in order to yield the hidden structures essential for curing and diagnosing diseases is frequently used. The EEG is therefore deemed a vital means for analysing brain function.

When electrical activity is recorded from the scalp, that recording contains electromyogram, or EMG, and the EMG is considered a serious contaminant of EEGs recorded from the scalp (Goncharova et al. 2003; McMenamin et al. 2010; Shackman et al. 2009). Stereotypically, EMG contamination is known to have large amplitude, which is why it is easily recognisable both visually and algorithmically. Moreover, it is generally the contaminated periods of EEG that are excised and discarded. However, constant weak contractions yield low amplitude impurities that are very stubborn in nature and difficult to detect visually. This continual contamination has spatial and spectral properties that are low power and difficult to recognise through the scalp recordings, but comparable to the contaminates cused by movement (Pope et al. 2009; Whitham et al. 2008). Temporary cranial, neck muscle and facial contractions result in electrical signals of very high amplitude with spectral features that overlap similar EEG bands. In addition, it has been established that recordings through the scalp and the range of incidences in the EMG interconnect, and as a result contaminate with the movement from muscles or EMG of the cranium and neck (Goncharova et al. 2003; Kumar et al. 2003).

The spatial resolution of the potential distributions is significantly reduced by the spatial smearing caused by the head volume conduction. For that reason, neck and face muscles have affected EEG signals recorded, and based on this, each electrode can be read for close and distant sources. Furthermore, surface Laplacian (SL) is sensitive to local sources as well as sources that are located close to the recording places and are impermeable to distant locations. Likewise, the SL diminishes enormously with the spatial smearing of the potential, which acts as a high-pass spatial filter (Nunez 1989). SL converts the existing scalp density with the help of data from all active scalp electrodes (Nunez & Srinivasan, 2006).

This paper will investigate the EMG contamination of the EEG signal. Raw data will be processed to clean it of EMG contamination by using ICA, and we will apply SL on the ICA data. Therefore, three kinds of data pre-processing will be used in artificial neural network (ANN) to classify neuropsychiatric diseases (anxiety, depression, and epilepsy) and control subjects based on the NBT features and for each type of pre-processing separately.

II. Hypothesis

The study goal is finding the effects of EMG contamination on the EEG signal with using classification to distinguish neuropsychiatric diseases (anxiety, depression and epilepsy) and control subjects. The pre-processing data that will be used are raw data, ICA (auto-pruned) data and ICA + SL. The study hypothesis is divided into three expected results, as shown in Table 1. The expected result (1) shows whether a difference in the data is caused by the muscles, so the brain activity has no differentiation between these tasks or diseases when applying muscle cleaning. For the second expected result (2), the brain has the same activity and muscles have no effect on brain activity, so all the results will be the same in each of the different data stages. In the expected result (3), the difference between these pre-processing types will increase with contaminated EMG. In this case, brain activity has been hidden by muscle contamination. Therefore, reading the EEG signal will be affected by the muscles. For example, we might expect that in the maze task there is more muscle contamination, so we would expect to see some like result 3 where the pre-processing methods reduce EMG contamination.

Table no 1: Expected result at three different data pre-processing stages.

Pre-processing data	Expected result (1)	Expected result (2)	Expected result (3)
Raw data	**Difference between tasks is higher than difference between them in ICA or ICA + SL.**	**Difference between tasks has not affected by muscles and has no different overall the data pre-processing.**	**Tasks has no different in this stage.**
ICA data	**Difference between tasks is higher than difference between them in ICA + SL.**		**Difference between tasks is higher than difference between them in Raw data.**
ICA + SL	**Tasks has no different in this stage.**		**Difference between tasks is higher than difference between them in raw data or ICA data.**

III. Methodology and Materials

EEG signals are usually used with neuropsychiatric diseases; therefore, this section examines the difference between those with neuropsychiatric diseases and control subjects. These diseases are anxiety, depression, and epilepsy. The study will compare each disease with controls under the three stages: raw data, data after applying ICA and data with combination of ICA and SL. In this section, the comparison will use machine learning to analyse data under NBT features. This section covers one of the main three expected results Table no 1.

Experimental subjects

This study uses data from subjects collected by The Brain Signals Lab (Whitham et al. 2007; Whitham et al. 2008). The subjects were chosen based on their diseases. Data was recorded with many tasks (Whitham et al. 2007; Whitham et al. 2008; DeLosAngeles 2010); however, eyes closed is the task that we chose for this study. The number of subjects in this study is 34, 10 were controls, 10 had depression, 10 had epilepsy and 4 had anxiety. Raw EEG signals were provided by The Brain Signals Lab. The Clinical Research Ethics Committee of the Flinders University and Flinders Medical Centre have given the approval for all experiments, and all subjects gave written informed consent (Fitzgibbon et al. 2016). All the data was recorded with 124 channels and 1000 Hz sample frequency. Data was prepared by applying ICA (auto-pruned method) on raw data and applying SL on data with ICA, which will be explained further later in this chapter.

Preparing the data

In this stage, this section has used the two stages of filtering to remove EMG contamination as used in the first section. The first filter is the ICA auto-pruned algorithm used to remove EMG contamination. The auto-pruned method uses AMICA for calculating the ICs that are used to prune the data. Then, the second filter is SL. We will use spherical spline SL to determine the local source of the electrode. As we have mentioned earlier, ICA isolates and removes EMG contamination; however, it may be affected by distant muscle sources, so SL collects the local sources of electrodes and rejects the distant sources. The combination of them isolates and removes the local and distant EMG contamination. In this section of the study, SL is applied to raw data as well to ensure the good results will only be affected by the SL or by the combination of ICA + SL.

The data was divided into one second segments because the samples were limited due the numbers of subjects with the studied diseases. Recording was done using 124 channels. Dividing data into one second segments will extend the data to be a large data set; therefore, machine learning will have a large data set for training and testing as shown in Table no2.

Features that will be used to examine the data are prepared by using NBT (https://www.nbtwiki.net/). NBT provides different kinds of computing biomarkers. The computing biomarkers that are used in this study are amplitude for some frequency bands (delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz)) and normalised amplitude for some frequency bands (delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz)). Each feature of these has used with 124 features that have given by the electrodes, therefore, each time of the classification has 124 features.

Table no 2: The number of actual subjects and the number of one second segments subjects for each disease and the control

	Anxiety	Depression	Epilepsy	Control
Actual number of subjects	**4 subjects**	**10 subjects**	**10 subjects**	**10 subjects**
Number of one second instances	**142 instances**	**360 instances**	**285 instances**	**348 instances**

Statistical analysis

Principal component analysis (PCA) is a method used for dimensionality reduction and feature extraction (Subasi &Gursoy 2010). PCA is used to represent the d-dimensional data in a lower-dimensional space that will minimise the degree of freedom and time complexities (Subasi &Gursoy 2010). Therefore, we have used PCA to reduce features, in some cases, to 9 features from 124 to get better and quicker results.

Ten-fold cross-validation is a method that is used to classify randomly split data to ten mutually exclusive subsets (the folds). Artificial neural network (ANN) is a MATLAB toolbox that performs a particular function of training a neural network by adjusting the values of the connection between elements (Demuth & Beale 1992). The subsets were entered into ANN to train the network using the Feed-Forward Neural Networks (FFNN) method (Bebis&Georgiopoulos 1994). This method works in one direction, which means there are no cycles or loops in the network (Zell 1994). FFNN has 1 hidden layer with 10 nodes. The algorithms used in this study are random data division, Levenberg-Marquardt to train the network, and Mean Squared Error in performance. Levenberg-Marquardt is an algorithm to solve the problem of minimising a non-linear function and is suitable for small and medium sized problems (Wilamowski& Yu 2010).

Study processing

The data used in this study was collected by The Brain Signal Lab (Whitham et al. 2007; Whitham et al. 2008; DeLosAngeles 2010) for the eyes closed task. Data is isolated and EMG contamination is removed by applying ICA, then by applying SL to remove distant muscle effects. Therefore, each kind of disease (anxiety, depression, and epilepsy) and the control data have four different kinds of data pre-processing: raw data, data with ICA, data with both ICA and SL, and raw data with SL. This data has been computed with the biomarkers (amplitude and normalised amplitude for different frequency bands (delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz)). The data has fewer subjects; therefore, we divide it into one second, non-overlapping segments to extend the data. PCA was applied to reduce the number of features, in some cases from 124 features to 9. Then, using the ten-fold cross-validation method, the data was entered into ANN to train the network with 1 hidden layer with 10 nodes and to find the generalisation accuracy percentage.

IV. Results and discussion

ANN was applied to classify the three neuropsychiatric diseases (anxiety, depression, and epilepsy) with control subjects under the four different types of data pre-processing (raw data, ICA data, ICA + SL, and raw data + SL) and with different features given by the NBT (https://www.nbtwiki.net/).

Table no 3 Accuracy percentages and biomarkers of classification of anxiety v control for each band with amplitude and normalised amplitude. The following symbols indicate significant differences: * from Raw, + from ICA, # from SL, ^ from ICA+SL

	Raw data		**ICA data**		**ICA + SL**		**SL**	
Frequency bands	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**
Amplitude								
Delta (1–4 Hz)	**72+#^**	**0.07**	**71*#^**	**0.01**	**96*+#**	**0.87**	**79*+^**	**0.32**
Theta (4–8 Hz)	**77#^**	**0.28**	**77#^**	**0.30**	**98*+#**	**0.96**	**89*+^**	**0.66**
Alpha (8–13 Hz)	**80+#^**	**0.39**	**82*#^**	**0.48**	**98*+#**	**0.93**	**95*+^**	**0.87**
Beta (13–30 Hz)	**89+#^**	**0.69**	**91*#^**	**0.37**	**99*+#**	**0.97**	**96*+^**	**0.90**
Gamma (30–45 Hz)	**92#^**	**0.79**	**93#^**	**0.80**	**98*+**	**0.97**	**98*+**	**0.94**
Normalised Amplitude								
Delta (1–4 Hz)	**71+#^**	**0.06**	**71*#^**	**0.67**	**92*+#**	**0.81**	**85*+^**	**0.63**
Theta (4–8 Hz)	**71#^**	**0.01**	**71#^**	**0.03**	**90*+#**	**0.75**	**81*+^**	**0.50**
Alpha (8–13 Hz)	**76+#^**	**0.28**	**75*#^**	**0.21**	**93*+#**	**0.79**	**87*+^**	**0.60**
Beta (13–30 Hz)	**75+#^**	**0.24**	**71*#^**	**0.06**	**94*+**	**0.88**	**95*+**	**0.87**
Gamma (30–45 Hz)	**96+#^**	**0.88**	**86*#^**	**0.61**	**100*+**	**0.99**	**100*+**	**1.00**

Anxiety versus control

Table no 3 shows the accuracy of classifying anxiety patients versus control subjects under the four different types of data pre-processing. The result shows no huge difference between raw and ICA data. The difference is usually 1%–2%. For example, the delta band in marked data gives higher accuracy (72%) than ICA data (71%) by 1%. Accuracy in the alpha band differed from ICA, which had higher accuracy (82%) than marked data (80%) by 2%. Also, for the gamma band, marked data had 92% accuracy in marked data and 93% in ICA data. On the other hand, the difference between SL and ICA + SL was obvious, especially in the delta and theta bands. However, the accuracy percentages were closer for the alpha and beta bands and similar in the gamma bands, which both had 98% accuracy. Table no 3 shows the obvious differences between the ICA + SL and both raw data and ICA in all frequency bands. Therefore, the good accuracy percentage for ICA + SL is based on both ICA + SL, even if ICA has not given a good result by itself.

Normalised amplitude gave a result quite similar to amplitude for the raw and ICA data, where there were no differences for the delta and theta bands and small differences between the alpha and beta bands. However, the gamma band has a huge difference in accuracy between them, where raw data has 96% accuracy

and ICA has 86%. For amplitude, ICA + SL has no differences in accuracy apart from in the beta band, where SL is 1% higher than ICA + SL.

In general, ICA + SL has given the best results in all bands, where the accuracy was greater than 95% for amplitude and greater than 90% for normalised amplitude. However, the best result was given by the gamma band for normalised amplitude for both ICA + SL and SL, which was 100% accuracy.

Table no 4 Accuracy percentages and biomarkers for classification of depression v control for each band for amplitude and normalised amplitude. The following symbols indicate significant differences: * from Raw, + from ICA, # from SL, ^ from ICA+SL.

	Raw data		**ICA data**		**ICA + SL**		**SL**	
Frequency bands	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**
Amplitude								
Delta (1–4 Hz)	**59+#^**	**0.18**	**62*#^**	**0.24**	**100*+**	**0.99**	**70*+**	**0.40**
Theta (4–8 Hz)	**66#^**	**0.33**	**65#^**	**0.30**	**98*+#**	**0.97**	**73*+^**	**0.47**
Alpha (8–13 Hz)	**55+#^**	**0.10**	**57*#^**	**0.15**	**100*+#**	**1.00**	**75*+^**	**0.50**
Beta (13–30 Hz)	**84#^**	**0.67**	**84#^**	**0.68**	**100*+#**	**1.00**	**90*+^**	**0.81**
Gamma (30–45 Hz)	**88#^**	**0.76**	**90#^**	**0.79**	**99*+#**	**0.99**	**94*+^**	**0.89**
Normalised Amplitude								
Delta (1–4 Hz)	**61#^**	**0.21**	**61#^**	**0.21**	**93*+#**	**0.85**	**65*+^**	**0.30**
Theta (4–8 Hz)	**55+^**	**0.11**	**57*^**	**0.14**	**92*+#**	**0.85**	**57^**	**0.14**
Alpha (8–13 Hz)	**56+#^**	**0.12**	**57*#^**	**0.13**	**92*+#**	**0.84**	**68*+^**	**0.36**
Beta (13–30 Hz)	**62#^**	**0.23**	**61#^**	**0.22**	**94*+#**	**0.87**	**72*+^**	**0.44**
Gamma (30–45 Hz)	**73#^**	**0.45**	**74#^**	**0.48**	**99*+#**	**0.97**	**88*+^**	**0.75**

Depression versus control

The result of classification of the depression patients and control subjects is shown in Table no 4. Amplitude features have shown small differences between marked and ICA data. For example, the delta band had 59% accuracy in the marked data and ICA 62%; for the theta band, marked data had 66% accuracy and ICA 65%; and marked data had 55% and ICA 57% in the alpha band, while there was improvement in accuracy in the gamma band between marked data and ICA data, from 88% to 90%. Moreover, SL data had better results than raw and ICA data, as shown in Table no 4; however, the ICA + SL gave the best result in all bands for amplitude. The delta, alpha and beta bands for amplitude gave 100% accuracy, and the gamma gave 99% accuracy.

The normalised amplitude results showed that the percentages are quite similar between the raw, ICA and SL data. For instance, the theta band in raw data gave 55%, whereas ICA and SL gave the same accuracy, 57%. The gamma band is the one where raw and ICA data gave large differences, with SL raw data achieving 73% accuracy and ICA 74%; whereas SL had 88%. Overall, ICA + SL gave the best result for normalised amplitude, where all bands had above 90% accuracy.

The gamma band for both amplitude and normalised amplitude gave 99% accuracy for ICA + SL data, as well as in this data the accuracy was similar or converged in other bands. For example, amplitude has three bands with the same 100% accuracy, and the rest approached 100%. Also, for normalised amplitude, the bands approached 93%, except the gamma band has greater accuracy than the others.

Table no 5 Accuracy percentages and biomarkers for classification of epilepsy v control for each band for amplitude and normalised amplitude. The following symbols indicate significant differences: * from Raw, + from ICA, # from SL, ^ from ICA+SL

	Raw data		ICA data		ICA + SL		SL	
Frequency bands	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**	**Accuracy %**	**BM**
Amplitude								
Delta (1–4 Hz)	**64+#^**	**0.28**	**67*^**	**0.16**	**84*+#**	**0.60**	**69*^**	**0.38**
Theta (4–8 Hz)	**66+#^**	**0.32**	**71*#^**	**0.34**	**83*+#**	**0.59**	**77*+^**	**0.54**
Alpha (8–13 Hz)	**64#^**	**0.28**	**66#^**	**0.16**	**82*+#**	**0.57**	**74*+^**	**0.48**
Beta (13–30 Hz)	**85#**	**0.70**	**85#**	**0.66**	**86#**	**0.63**	**82*+^**	**0.65**
Gamma (30–45 Hz)	**93+#^**	**0.86**	**92*^**	**0.82**	**96*+#**	**0.91**	**92*^**	**0.84**
Normalised Amplitude								
Delta (1–4 Hz)	**60+#^**	**0.19**	**64*#^**	**0.12**	**77*+#**	**0.47**	**68*+^**	**0.36**
Theta (4–8 Hz)	**62#^**	**0.25**	**62#^**	**0.05**	**66*+#**	**0.19**	**67*+^**	**0.35**
Alpha (8–13 Hz)	**59+#^**	**0.17**	**64*#^**	**0.12**	**77*+#**	**0.46**	**67*+^**	**0.34**
Beta (13–30 Hz)	**70#^**	**0.39**	**71#^**	**0.28**	**80*+#**	**0.55**	**74*+^**	**0.48**
Gamma (30–45 Hz)	**82+#^**	**0.64**	**76*#^**	**0.43**	**87*+#**	**0.69**	**91*+^**	**0.82**

Epilepsy versus control

For this classification, the reduction in accuracy of all results was apparent when compared with the other classifications. Moreover, the accuracy percentages for the delta to gamma bands do not differ from those of the other classifications, as shown in Table no 5. For example, raw data in the delta band has 64% accuracy, and gamma has 93%. However, the alpha band for each type of pre-processing for amplitude is less accurate than the theta band, which did not occur for the other classifications (Tables 3 and 4). For instance, for raw data, the theta band has 66% accuracy, and alpha has 64%; for ICA data, the theta band has 1% accuracy, and alpha has 66%. For amplitude at all frequency bands, ICA + SL gave the best result of all data pre-processing. The gamma band with ICA + SL gave 96% accuracy, the highest accuracy of all bands.

The disparity between pre-processing is not great, especially between raw, ICA and SL data. For example, the delta band raw data got 64% accuracy, ICA 67%, and SL 69%. While the disparity between them and ICA + SL is obvious in the lower bands, it is not as great in the higher bands. For instance, the delta band ICA + SL had 84% accuracy, which is great in comparison with the others; however, the beta band ICA+ SL had 86%, while raw data and ICA data had 85% and SL had 82%.

Normalised amplitude had different results from amplitude, with disparities in accuracy between the bands for each type of pre-processing. For example, raw data for the alpha band had 59% accuracy, while delta had 60%, and theta had 62%. Also, for ICA and ICA + SL, delta and alpha have the same accuracy percentages, while theta is less accurate. SL gave the highest accuracy in the gamma band, where it was 91%. The gamma band ICA + SL was less accurate than SL, which is due to the disparity between raw data and ICA data, where raw data had 82% while ICA data had 76%.

T-test

Student's t-test has been used for statistically analysing the results. The t-test was calculated for each band in both amplitude and normalised amplitude frequency bands between the pre-processing data. Tables 3, 4 and 5 show the significant differences and non-significant differences between the data pre-processing types for each classification ($p < 0.05$).

In delta and alpha bands over both amplitude and normalised amplitude usually give significant different level between data pre-processing. However, the other bands have different result from one classification table to other table.

The t-test results for raw and ICA data shows non-significant difference in more than one of the different frequency bands. Most of the time, the non-significant difference arose between those data pre-processing in all classification tables and over all bands, were 13 out of 30. ICA + SL has significant difference

with each pre-processing over all bands in each Tables 3, 4 and 5. ICA+SL has proved that the combination between those pre-processing gives the best result overall all bands.

As mentioned previously, the SL has used to confirm that the ICA+SL is affected only by influence of SL or by the combination of both methods. The differences in the accuracy percentages have shown that as well as the t-test with the significant different in the almost all the t-test between ICA+SL and SL data pre-processing. Therefore, the ICA+SL is an effective combination of both methods.

EMG contamination

Classification of diseases under the pre-processing data gave different accuracies, shown in Tables 3, 4 and 5. ICA data has non-significant differences with raw data more than other data pre-processing, which means ICA did not quite improve data, similar to in the first section. In this case, there may be two reasons for that. The first is the classification was performed on 124 channels on the scalp, and some have minimal muscle contamination (Fitzgibbon et al. 2016). Accuracy percentages for raw data and SL in Tables 3, 4 and 5 show small improvements over raw data and significant different in t-test in the most bands. Therefore, we can say that combination of ICA + SL improved both t-test and accuracy. As we mentioned in the first section, ICA is able to isolate and remove the EMG contamination and SL collects data from local sources. These features in the combination of ICA and SL proved the first reason. The second reason is the number of subjects in the study was limited. The number of subjects for training and testing the validation was limited, which may have affected identification of the features that were hidden by EMG contamination. SL makes the features that were hidden by EMG clear; hence, the best result was from ICA + SL.

V. Conclusion

This paper has demonstrated the effect of EMG on the EEG signal by comparing EEG signals under three different types of data pre-processing. The study was used four types of data pre-processing: raw data (no pre-processing), data after applying ICA, data after applying ICA + SL, and raw data + SL.

With using machine learning to classify neuropsychiatric diseases (anxiety, depression, and epilepsy) and control subjects under the four types of data pre-processing (raw data, ICA, ICA + SL, and SL). ANN was used for training data and testing validation. The features were extracted from NBT, which were amplitude and normalised amplitude for all frequency bands. Also, the Student's t-test was applied to discoverthe significant differences and non-significant differences between types of pre-processing for amplitude and normalised amplitude for all bands. The result was that SL had the highest accuracy for all the bands and had significant differences between it and raw data for anxiety v control and depression v control, and non-significant differences for epilepsy v control, with obvious differences in accuracy percentages in all bands. However, ICA had non-significant differences for all the classifications with raw data in the t-test and showed no improvement in accuracy percentages. Moreover, SL gave non-significant differences in the t-test with raw data; however, with the observed bands, accuracy percentages are improved.

In general, the result was between the second and third expectations. ICA does not improve the accuracy percentages, which means the EMG contamination did not affect brain activity for the classification. However, ICA + SL improved the accuracy percentages, which means EMG contamination affects brain activity and by removing EMG contamination, the accuracy was improved. The effect of SL was not the only reason for the improvement in the accuracy, which was confirmed when we applied SL to raw data giving small improvements in accuracy. Therefore, ICA played role in improving the results when integrated with SL.

Study limitations

The NBT that we used is version 0.5.5-public, which has limitations in that some features cannot gives limitation in result whatever the data that has been computed. For example, Coherence, Phase Locking Value, phase looking value and Detrended fluctuation analysis (DFA) Also, for biomarker statistics we had to use the MATLAB version 2014a to display the figures. As well as the statistical tests some of them have not display figures such as one-way or two-way ANOVA, Wilcoxon paired sum test and Permutation test for paired mean difference.

The data set had a small number of subjects for training the ANN and testing the validation, which may have affected the results.

This study used 124 channels to examine the entire scalp. Some of these channels are affected by EMG contamination, and some diseases are different from normal in specific regions of the brain while the rest has the same brain activity; therefore, we believe that has affected the results, especially for classification.

Another factor to consider is that subjects were intermixed between training and testing data/fold (i.e. it did not follow a leave one subject out methodology) and so the overall accuracies may be inflated. However, this is consistent across all pre-processing methods and so valid conclusions can still be drawn between the methods.

Future work

This study has used 124 channels from all the brain regions. However, in future work, the classification of neuropsychiatric diseases and control subjects must be specific on the regions of differentiation between each disease and the controls. As well, the number of subjects must be increased to give more accurate results.

Amplitude and normalised amplitude are the features that have been used in this study. However, it would be interesting to investigate further features such as bandwidth (BW), peak frequency, spectral edge frequency (SEF), root mean-squared EEG amplitude (RMS Amp), minima and maxima, and Shannon entropy (HSH).

The focus of this paper was the different muscle reducing pre-processing methods and not necessarily the machine learning algorithms. It would be interesting to investigate further using the dataset with different machine learning algorithms such as SVM or even Deep Learning if the data is sufficiently large.

References

[1]. BASHASHATI, A, FATOURECHI, M, WARD, RK & BIRCH, GE 2007, 'A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals', Journal of Neural Engineering, vol. 4, no. 2, R32-57.

[2]. BASHASHATI, A, FATOURECHI, M, WARD, RK & BIRCH, GE 2007, 'A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals', Journal of Neural Engineering, vol. 4, no. 2, R32-57.

[3]. BEBIS, G & GEORGIOPOULOS, M 1994, 'Feed-forward neural networks', IEEE Potentials, vol. 13, no. 4, pp. 27-31.

[4]. BELL, AJ & SEJNOWSKI, TJ 1995, 'An information-maximization approach to blind separation and blind deconvolution', Neural Computation, vol. 7, no. 6, pp. 1129-1159.

[5]. DELOSANGELES, D 2010, Electroencephalographic, Cognitive and Autonomic Correlates of States of Concentrative Meditation, Flinders University, Adelaide.

[6]. DEMUTH, H & BEALE, M 1992, Neural network toolbox. For Use with MATLAB. The MathWorks, Inc, Natick.

[7]. FITZGIBBON, S, DELOSANGELES, D, LEWIS, T, POWERS, D, GRUMMETT, T, WHITHAM, E, WARD, L, WILLOUGHBY, J & POPE, K 2016, 'Automatic determination of EMG-contaminated components and validation of independent component analysis using EEG during pharmacologic paralysis', Clinical Neurophysiology, vol. 127, no. 3, pp. 1781-1793.

[8]. GARRETT, D, PETERSON, DA, ANDERSON, CW & THAUT, MH 2003, 'Comparison of linear, nonlinear, and feature selection methods for EEG signal classification', IEEE Transactions on Neural Nystems and Rehabilitation Engineering, vol. 11, no. 2, pp. 141-144.

[9]. GONCHAROVA, II, MCFARLAND, DJ, VAUGHAN, TM & WOLPAW, JR 2003, 'EMG contamination of EEG: spectral and topographical characteristics', Clinical Neurophysiology, vol. 114, no. 9, pp. 1580-1593.

[10]. KACHENOURA, A, ALBERA, L, SENHADJI, L & COMON, P 2008, 'ICA: a potential tool for BCI systems', IEEE Signal Processing Magazine, vol. 25, no. 1, pp. 57-68.

[11]. KUMAR, S, NARAYAN, Y & AMELL, T 2003, 'Power spectra of sternocleidomastoids, splenius capitis, and upper trapezius in oblique exertions', The Spine Journal, vol. 3, no. 5, pp. 339-350.

[12]. LEE, T-W, LEWICKI, MS, GIROLAMI, M & SEJNOWSKI, TJ 1999, 'Blind source separation of more sources than mixtures using overcomplete representations', IEEE Signal Processing Letters, vol. 6, no. 4, pp. 87-90.

[13]. MCMENAMIN, BW, SHACKMAN, AJ, MAXWELL, JS, BACHHUBER, DR, KOPPENHAVER, AM, GREISCHAR, LL & DAVIDSON, RJ 2010, 'Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG', Neuroimage, vol. 49, no. 3, pp. 2416-2432.

[14]. NUNEZ, PL 1989, 'Estimation of large scale neocortical source activity with EEG surface Laplacians', Brain Topography, vol. 2, nos. 1-2, pp. 141-154.

[15]. NUNEZ, PL & SRINIVASAN, R 2006, Electric fields of the brain: the neurophysics of EEG, Oxford University Press, USA.

[16]. Oja, H. and Nordhausen, K., 2001. Independent component analysis. Encyclopedia of Environmetrics.

[17]. POPE, KJ, FITZGIBBON, SP, LEWIS, TW, WHITHAM, EM & WILLOUGHBY, JO 2009, 'Relation of gamma oscillations in scalp recordings to muscular activity', Brain topography, vol. 22, no. 1, pp. 13-17.

[18]. SANNELLI, C, BRAUN, M & MÜLLER, K-R 2009, 'Improving BCI performance by task-related trial pruning', Neural Networks, vol. 22, no. (9), pp. 1295-1304.

[19]. SHACKMAN, AJ, MCMENAMIN, BW, SLAGTER, HA, MAXWELL, JS, GREISCHAR, LL & DAVIDSON, RJ 2009, 'Electromyogenic artifacts and electroencephalographic inferences', Brain Topography, vol. 22, no. 1, pp. 7-12.

[20]. SUBASI, A & GURSOY, MI 2010, 'EEG signal classification using PCA, ICA, LDA and support vector machines', Expert Systems with Applications, vol. 37, pp. 8659-8666.

[21]. WHITHAM, EM, LEWIS, T, POPE, KJ, FITZGIBBON, SP, CLARK, CR, LOVELESS, S, DELOSANGELES, D, WALLACE, AK, BROBERG, M & WILLOUGHBY, JO 2008, 'Thinking activates EMG in scalp electrical recordings', Clinical neurophysiology, vol. 119, no. 5, pp. 1166-1175.

[22]. WHITHAM, EM, POPE, KJ, FITZGIBBON, SP, LEWIS, T, CLARK, CR, LOVELESS, S, BROBERG, M, WALLACE, A, DELOSANGELES, D & LILLIE, P 2007, 'Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG', Clinical Neurophysiology, vol. 118, no. 8, pp. 1877-1888.

[23]. WILAMOWSKI, BM & YU, H 2010, 'Improved computation for Levenberg–Marquardt training', IEEE Transactions on Neural Networks, vol. 21, no. 6, pp. 930-937.

[24]. WOLPAW, JR, BIRBAUMER, N, HEETDERKS, WJ, MCFARLAND, DJ, PECKHAM, PH, SCHALK, G, DONCHIN, E, QUATRANO, LA, ROBINSON, CJ & VAUGHAN, TM 2000, 'Brain-computer interface technology: a review of the first international meeting', IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pp. 164-173.

[25]. ZELL, A 1994, Simulation neuronalernetze, Addison-Wesley, Bonn.