AUC Confidence Bounds for Performance Evaluation of Brain-Computer Interface

Brahim Hamadicharef
Tiara #22-02, 1 Kim Seng Walk, Singapore 239403
bhamadicharef@hotmail.com

***Abstract*—Currently most performance evaluation of Brain-Computer Interface (BCI) systems is simply reported in terms of accuracy. In this paper we propose a novel approach to evaluate the true performance of BCI systems based on Receiver Operating Characteristic (ROC) analysis, that removes the limitations of the accuracy performance measure. We demonstrate the need to provide, and particularly for small sample size, Confidence Interval (CI) bounds to indicate reliability of the BCI system performance. The ROC-based methodology makes it possible to calculate CI, shown as a contour at each any points of the ROC curve, with value of the lower bound of the Area Under the Curve (AUC). We illustrate the usefulness of the methodology using the results the BCI Competition IV data set 3, dealing with the classification of wrist movements from four directions recorded using magnetoencephalogram (MEG). Plotting the 95% CI contours overlayed on the ROC curves revealed some overlap with the chance level, thus revealing potential different interpretation from claims based on single accuracy value. The ROC-based methodology will also help to determine minimal sample size, an important requirement for future BCI studies and competitions.**

I. Introduction

Over the last two decades Brain-Computer Interface (BCI) systems have gained a great deal of popularity, with a sign is its growing number of publications and regular organization of world competitions during which researchers are proposing innovative solutions to the BCI challenge. Despite a tool-set of methods available for evaluating the performance of BCI systems [1], a simple figure of accuracy is typically used. No robust performance evaluation methodology has yet been put forward.

This comes mainly from the fact that the number of trials recorded differs from studies to studies, some results are provided with accuracy while others provide a more elaborated performance evaluation e.g. 10x10 fold cross-validation. This issue is even more appropriate when looking at the recent interest in BCI which include zero-training aspects [2], as researchers are trying to minimize the generality of their algorithms.

Rigorous evaluation should be of primary concern if the BCI research community seeks to fully exploit the potential of BCI systems. Concerns should be taken to avoid cases such as the one recently published in [3], in which authors discussed the potentially misleading technological aspects of a commercialized BCI headset. Another recent article published promising results on a Near Infrared Spectroscopy (NIRS) based system claiming high 80% accuracy, but with a weak methodology as found out by Dominguez [4]. This was followed by a reply [5] from the original authors admitting much lower accuracy results (average 53%, 3 out of 9 with only 63% accuracy). Such cases should serve as warnings for the BCI research community to seek more rigorous and robust methodologies for performance evaluation.

From the literature, it has been shown that AUC should be preferred over accuracy [6]. Other research domains, in particular related to medicine such as radiology, consider that reporting only accuracy is not enough and have imposed AUC as the de-facto measure of performance in particular with small sample size [7].

In this paper we contribute by proposing a methodology based on Receiver Operating characteristic (ROC) [8] analysis that makes use of Probability Density Function (PDF) to calculate the confidence interval (CI) shown as contours overlayed at any ROC point, and using a fast version [9] both lower and upper CI bounds of the Area Under the Curve (AUC). The methodology based ROC analysis was first developed for objective evaluation of intelligent medical systems [10]. From its constituent parts, a BCI system makes use of signal processing, for the Electroencephalogram (EEG) pre-processing and feature extraction and machine learning stage, with a typical classification based on Linear Discriminant Analysis (LDA), and can thus be considered as an intelligent medical system.

One particular limitation of current BCI research is that most studies report small sample size not only in terms of number of trials used to train and test BCI systems, but also in terms of subjects recruited/enrolled in BCI studies. This should motivate to use more robust, statistically sound methodologies, such as ROC analysis to evaluate BCI systems.

Finally, unlike other medical systems for which the cost of diagnosis is important (e.g. cancer or brain disease), BCI systems currently aim at relatively simple control of a wheelchair [11] and communication applications such as P300-based speller, thus even if the safety aspect in BCI is still relatively minimal, it remains paramount for future clinical and home usage. In this respect, ROC analysis also serves better the performance evaluation of BCI systems.

The remainder of this paper is organized as follows. In Section II we propose a methodology based on ROC analysis. We illustrative its use with BCI example in Section III. Finally, in Section IV, we conclude the paper.

978-1-4244-6497-5/10/$26.00 ©2010 IEEE

TABLE I
CONTINGENCY TABLE

	Gold Standard	
Test	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

II. ROC-BASED METHODOLOGY

A typical classifier, used in BCI systems, aims to separate two or more classes using a set of features extracted from the data. For non-invasive BCI systems, such data are typically MEG, EEG or NIRS time segments, called trials, associated to a specific tasks such as real hand movements [12], finger tapping, motor imagery, etc. A 2-class classifier can have four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) making a contingency table (or confusion matrix) as shown in Table I.

From a $\{$TP,TN,FN,FN$\}$ set one can calculate performance measures such as sensitivity (SEN = $\frac{TP}{TP+FN}$), specificity (SPE = $\frac{TN}{TN+FP}$) and accuracy (ACC = $\frac{TP+TN}{TP+TN+FP+FN}$). Using a threshold that varies across the full range of the classifier's output, one can obtain a multitude of contingency tables and plot of the Sensitivity versus 1-Specificity which is known as a Receiver Operating Characteristic (ROC) curve [8].

One of the main limitations of accuracy can be demonstrated from the following. If we take one of the BCI Competition IV result, the winning solution for example, the accuracy for Subject1 is 59.5%. Such figure corresponds to a ratio of 44/74. If another BCI system is developed and based on a larger number of trials, say 370, the a 88/370 will have the same accuracy. However, from a machine learning point of view, the later system seems 'better' than the first one, because its accuracy is calculated over a larger number of trials, thus improving the effect of generalization. Such simple example shows that to provide only an accuracy measure for performance is not enough and that only CI allows to define some reliability of the measure. This should be even more important when dealing with small size data sets as it is often the case in BCI.

We thus propose to use a ROC-based methodology that will remove these current limitations. Initially developed to evaluate intelligent medical systems [10][13], it is adapted to BCI systems. Due to the small sample size, either in terms of number of trials or subjects in BCI studies, this methodology is best suitable for the task of evaluating the performance BCI systems. It was also shown to be able to estimate sample size requirements [14].

III. ILLUSTRATIVE EXAMPLE

To illustrate our approach we use the results from the BCI Competition IV data sets 3[1] on hand movement [12]. The true labels were provided post-competition for both Subject1 and Subject2. As shown in Figure 1, one should notice that Subject2 has one less test trial than Subject1, and both have unbalanced class label count (four directions with 14-30-15-15 for Subject1, and 19-18-12-24 for Subject2).

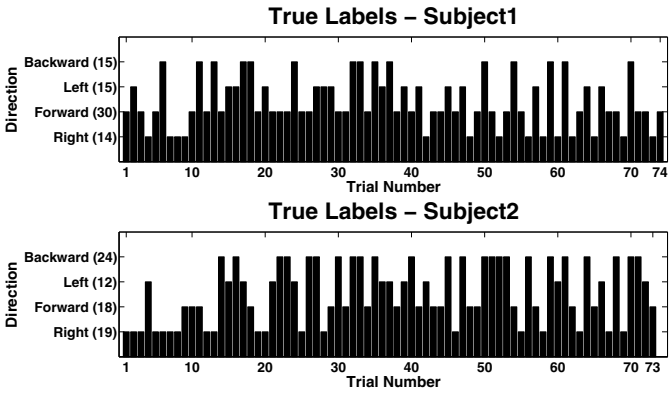


Fig. 1. BCI Competition IV data sets 3: true labels

The four winner solutions, from Team1 to Team4, are as follows. First place was given to S. Hajipour and M. B. Shamsollahi (Sharif University of Technology, Tehran, Iran) with an average accuracy of 46.9% (Subject1: 59.5%, Subject2: 34.3%). Second place was attributed to J. Li with W. Hong, J. Song, Y. Xu, X. Li (Yanshan University, Qinhuangdao, China), with an average accuracy of 25.1% (Subject1: 31.1% and Subject2: 19.2%) Third place was for N. Montazeri and M. B. Shamsollahi (Sharif University of Technology, Tehran, Iran) with an average accuracy of 23.9% (Subject1: 16.2% and Subject2: 31.5%), while fourth place to J. Wang and T. Zhang (Yanshan University, Qinhuangdao, China) with an average accuracy 20.4% (Subject1: 23.0% and Subject2: 17.8%).

For a 4-class classification paradigm, the level of chance is at the accuracy of 25%, thus all results from the Team4 are not better than chance. To be complete, a 4-class classification paradigm chance level is not exactly 25%, one should also consider the CI with a level α which depends on the number of trials. For example, the upper level is calculated as 29.7% for 80 trials and α = 5% i.e. 95% CI (See [15] for details).

Using known accuracy results and number of trials from the true labels (i.e. N_{Trials} = TN+TP+FN+FP), we enumerate all possible cases of TN+TP using the accuracy ratio. We can thus plot ROC curve corresponding to this specified accuracy (by enumerating TN and TP, we find specificity and sensitivity, accordingly). We show such ROC curves for each team in Figure 2, with the chance level (*line of chance*) at 50% for 2-class paradigm and 25% for 4-class paradigm.

As shown in Figure 2(a), of all submissions, only the winning team achieved both accuracy for Subject1 and Subject2 to be higher than chance level. Team2 and Team3 managed to get better than chance for Subject1 (31.1%) and Subject2 (31.5%), respectively, as shown in Figure 2(b) and Figure 2(c). However these results are rather close to the 29.7% for the chance level (upper 95% CI).

[1] http://www.bbci.de/competition/iv/#dataset3

To illustrate the use of the ROC methodology, we plot few contours for specific ROC points in Figure 2(a), Figure 2(b) and Figure 2(c). For Subject1, N_{Trials} is 74 and for Subject2 N_{Trials} is 73, while for the mean ROC curve, we concatenate trials from both subjects' and thus used N_{Trials} to be 148 (i.e. 74+73).

It is rather obvious from the figures that, because of the small sample size, the 95% CI contours are rather large. As shown in Figure 2(a), for Team1 both contours for Mean (AUC-L=0.371) and Subject1 (AUC-L=0.455) AUC-L 95%CI are well above the chance level (AUC=0.25), however not for Subject2 (AUC-L=0.220), where there is clear overlap. For the all 3 other teams, CI contours all overlap the level of chance, providing little confidence about the performance due to small sample size. One should also notice that because of the larger number of trials N_{Trials} (i.e. 74+73, as we concatenate Subject1 and Subject2 results), the 95% CI contour for the mean is smaller than the one of Subject1 or Subject2, as expected in Figure 2(a) and Figure 2(d).

In [14], the ROC-based methodology was found useful for sample size determination (SSD), i.e. finding the minimum number of trials, from a ROC analysis stand-point, that will allow adequate statistical validity for performance. Results, useful for future BCI studies and BCI competitions, will be presented elsewhere.

IV. Conclusions

In this paper we looked at performance evaluation for BCI systems, one important issue and in particular with small sample size. We provide a critical analysis of the current limitations of accuracy and proposed an approach, based on ROC analysis with CIs, provide confidence (e.g. at 95% or 99%) in BCI systems accuracy results typically reported in the BCI literature.

Illustrative example on the BCI Competition IV data sets 3, provided a real of lack of confidence about the current performance presented as accuracy. We argue that more adequate measures, such as ROC curve, AUC, which should be, as they are based on robust statistical foundations, be important part of the BCI research tool-set.

This study helped to rise two important conclusions. First, to improve the current state-of-the-art, BCI systems performance evaluation should be reported with figures of merit such as the $\{$sensitivity, specificity, accuracy$\}$ triplet, p-value from statistical significance test, and as proposed here a ROC analysis from which a ROC curve with 95% CI contours can be visualized, calculate the AUC with CI (lower and upper bounds as derived in [14][9]) and Standard Error (SE) [16]. Alltogether they form a comprehensive and robust performance evaluation for BCI systems.

Finally, as a young research field, we believe that there is a need to provide adequate and unified guidelines to report BCI results, to avoid methodological mistakes [4] and associated negative publicity. This will only be achieved by mean of reproducible results supported by robust evaluation, e.g. based on ROC analysis. We believe that the BCI research community should also seek to elaborate *guidelines for reporting BCI studies*, as in other domains e.g. in neuroimaging [17], or in the form of simple rules [18]. Finally, we should also seek to use large data set for performance evaluation such as the recent initiative by the Team PhyPA (Physiological Parameters for Adaptation)[19].

References

[1] A. Schlögl, J. Kronegg, J. E. Huggins, and S. G. Mason, "Evaluation criteria in BCI research," *In: G. Dornhege, J. del R. Millán, T. Hinterberger, D. J. McFarland, K.-R. Müller (Eds.). Toward brain–computer interfacing, MIT Press*, pp. 327–342, 2007.

[2] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards Zero Training for Brain–Computer Interfacing," *PLoS ONE*, vol. 3, no. 8, 2008.

[3] D. Heingartner, "Mental Block," *IEEE Spectrum*, vol. 46, no. 1, pp. 34–35, January 2009.

[4] L. G. Dominguez, "On the risk of extracting relevant information from random data," *Journal of Neural Engineering*, p. 2pp, 2009, 058001.

[5] T. Chau and S. Damouras, "Reply to 'On the risk of extracting relevant information from random data'," *Journal of Neural Engineering*, vol. 6, 2009, 058002.

[6] C. Ling, J. Huang, and H. Zhang, "AUC: a Statistically Consistent and more Discriminating Measure than Accuracy," *Proceedings of the 18th International Joint Conference on Artiicial Intelligence (IJCAI-2003), Acapulco, Mexico, August 9-15, 2003*, 2003.

[7] N. A. Obuchowski and M. L. Lieber, "Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples," *Academic Radiology*, vol. 5, no. 8, pp. 561–571, August 1998.

[8] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, October 1978.

[9] B. Hamadicharef, "Frequentist versus Bayesian approaches for AUC Confidence Interval Bounds," *Proceedings of the 10th International Conference on Information Science, Signal Processing and their applications (ISSPA2010), Kuala Lumpur, Malaysia, May 10–13, 2010*, pp. 341–344.

[10] J. Tilbury, P. Van-Eetvelt, J. Garibaldi, J. Curnow, and E. Ifeachor, "Receiver Operator Characteristic Analysis for Intelligent Medical Systems - A New Approach for Finding Confidence Intervals," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 952–963, July 2000.

[11] B. Rebsamen, E. Burdet, C. Guan, H. Zhang, C. L. Teo, Q. Zeng, C. Laugier, and A. J. M. H, "Controlling a Wheelchair Indoors Using Thought," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 18–24, March/April 2007.

[12] S. Waldert, H. Preissl, E. Demandt, C. Braun, N. Birbaumer, A. Aertsen, and C. Mehring, "Hand Movement Direction Decoded from MEG and EEG," *The Journal of Neuroscience*, vol. 28, no. 4, pp. 1000–1008, January 2008.

[13] E. C. Ifeachor and B. Hamadicharef, "ROC Analysis in the Evaluation of Intelligent Medical Systems," *Proceeding of the 1st European Workshop on the Assessment of Diagnostic Performance (EWADP'2004), Milan, Italy, July 7–9, 2004*, pp. 23–32.

[14] V. Stalbovskaya, B. Hamadicharef, and E. C. Ifeachor, "Sample Size Determination using ROC Analysis," *Proceeding of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), Plymouth, U.K., July 25–27, 2007*.

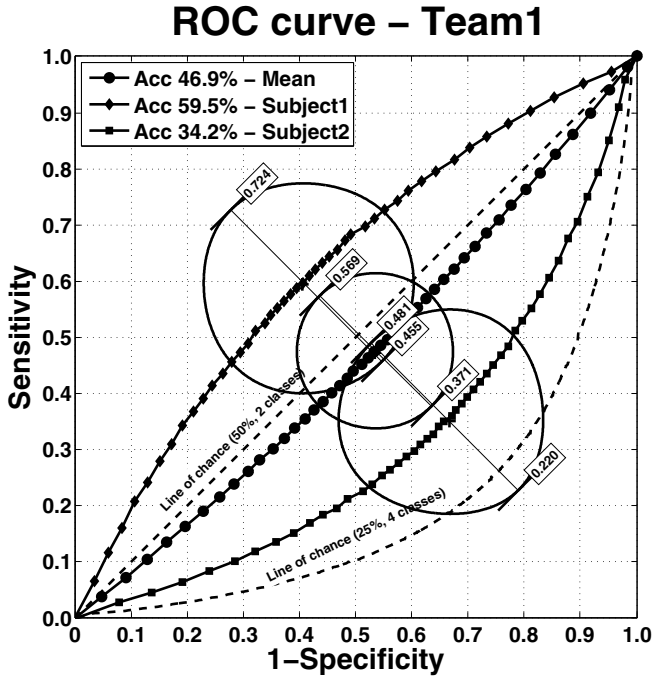
[15] G. R. Müller–Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random? A closer look on BCI results," *International Journal of Bioelectromagnetism*, vol. 10, no. 1, pp. 52–55, 2008.

[16] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

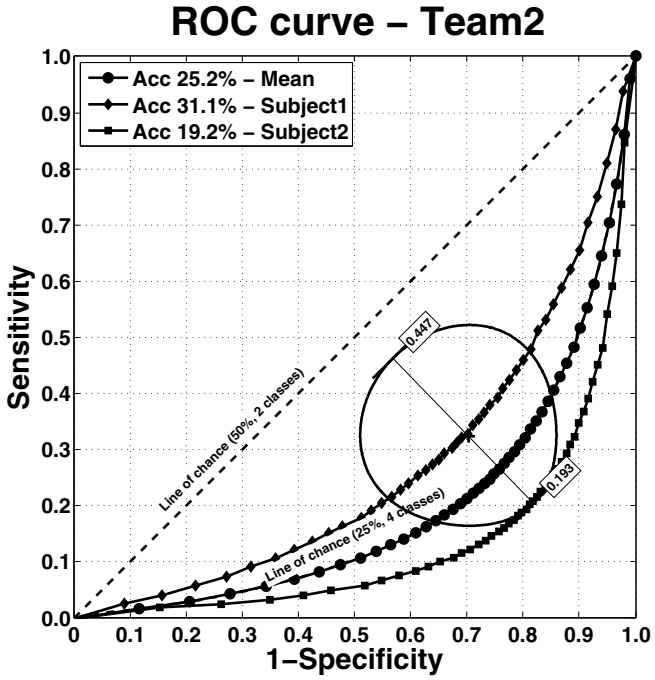
[17] R. A. Poldrack, P. C. Fletcher, R. N. Henson, K. J. Worsley, M. Brett, and T. E. Nichols, "Guidelines for reporting an fMRI study," *NeuroImage*, vol. 40, no. 2, pp. 409–414, April 2008.

[18] G. R. Ridgway, S. M. D. Henley, J. D. Rohrer, R. I. Scahill, J. D. Warren, and N. C. Fox, "Ten simple rules for reporting voxel–based morphometry studies," *NeuroImage*, vol. 40, no. 4, pp. 1429–1435, 2008.

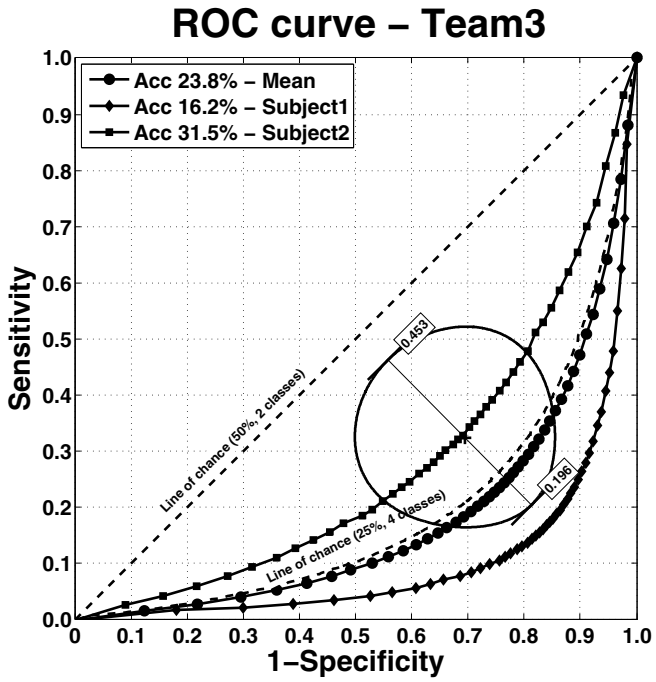
[19] T. Zander and C. Kothe, "PhyPA Benchmarking common BCI algorithms," *(Technische Universitaet Berlin, Department for Human-Machine Systems, Team PhyPA)*, 2008, http://www.phypa.org.



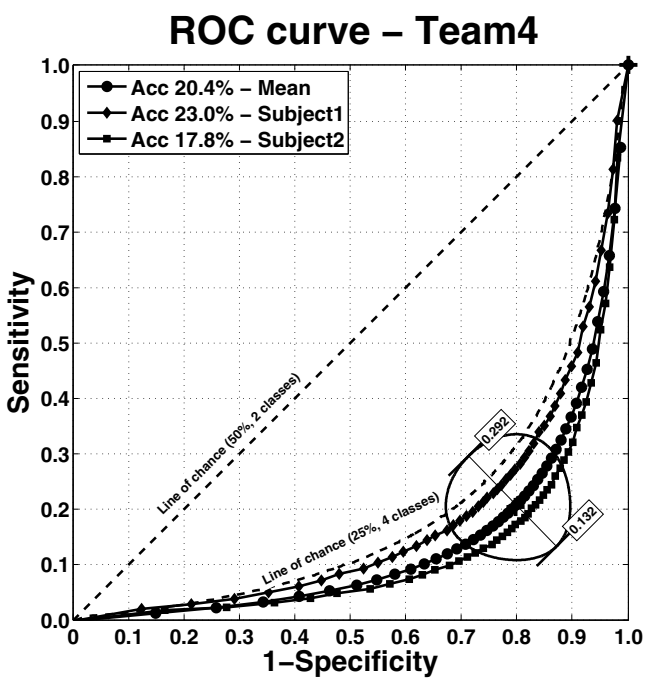
(a) 1st place



(b) 2nd place



(c) 3rd place



(d) 4th place

Fig. 2. ROC curve with 95% CI for each results of BCI Competition IV data sets 3