JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 8(3), SEPTEMBER, 2020
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

Measures and Metrics for Assessing Detection Algorithms in Brain Machine Interfaces

[1]Ameer Mohammed, [2]Ashraf A. Ahmad, [3]Auwalu M. Abdullahi, [4]Mahmood T. Kabir
[1]Mechatronic Engineering Department, Air force Institute of Technology, Kaduna.
[2]Department of Electrical/Electronic Engineering, Nigerian Defence Academy, Kaduna.
[3]Department of Mechatronics Engineering, Bayero University, Kano.
Department of Electronics and Telecommunication Engineering, A. B. U., Zaria.

ABSTRACT
A major challenge in the analysis of detection algorithms for brain machine interface (BMI) is the availability of objective methods for comparing various technologies. The absence of standard measures for BMI detection algorithms prompted the need for a thorough review and commentary on performance measures for BMI algorithms. This was necessary so that suitably tailored measures are selected that are fully representative and compatible with set performance targets. The review includes measures that could be used to evaluate efficacy, complexity and efficiency of detection algorithms; as well as their strengths and weaknesses. At the end, a commentary is provided on the need to provide either standard or custom performance measures for BMI algorithms, such that the algorithms can be assessed accurately.

ARTICLE INFO
Article History
Received: February, 2020
Received in revised form: March, 2020
Accepted: June, 2020
Published online: August, 2020

KEYWORDS
Bio-signal processing; Brain machine interface; Deep brain stimulation; Feedback algorithms; Machine learning; Neural prosthesis; Performance measures and SIC,

INTRODUCTION

Research in BMI has come a long way, so far there have been seven international BMI meetings, in which the most recent took place in 2018. As the field progresses, there is a need to provide objective methods to compare BMI technologies. It is for this reason studies like (Mason & Birch, 2003) have attempted to provide *experimental frameworks* on how to compare BMI. However, a very conspicuous gap has been the lack of standard metrics for assessing BMI systems. With advancements in the use of BMI for neural rehabilitation, the rise in the significance of detection algorithms as their major building blocks cannot be overemphasised. In BMI applications, detection algorithms convert recorded brain activity into useful information that could be translated or used as the control signals for the control of sensor motor functions. Detection algorithms are the first step towards developing BMIs that uses pathological brain activity to ameliorate, mitigate or restore bodily function in patients with disabilities such as such as Parkinson's disease, Alzheimer's disease, epilepsy, spinal cord injury (SCI) and so on. In addition, BMIs have recently been found to be useful in the restoration of lost emotional function in neuropsychiatric disorders by activating the neural mechanisms responsible for the regulation of emotions (Shanechi, 2019).That is why BMIs have been the major cornerstone of the BRAIN initiative (Brain Research through Advancing Innovative Neuro-technologies), which is a $100 million investment in research and development with the goal of supporting the development and application of innovative technologies that can create a dynamic understanding of brain function (Secretary, 2013).

In BMI, efficient and efficacious detection algorithms could lead to fully implantable systems with real-time capabilities that could be used to monitor

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

the functionality of prosthesis or disease progression in patients. Also, the economic impact is another motivation for understanding the various metrics that can be used in evaluating detection algorithms. Currently, there is an increase in cost of management for neuro-motor disabilities both for individuals (Marceglia et al., 2015) and health care administrators (French et al., 2007). As such, it is paramount to review and analyse various performance metrics in BMI detection algorithm as they influence this cost of BMI detection algorithms. This will enable proper resource utilisation for hardware efficient implementation. In this work, the focus will be only BMI detection algorithms. The metrics were categorised into three, based on what characteristic they evaluate. The three categories are: complexity, efficacy and efficiency. Of these three categories of measures, there are more subtle differences between efficacy and efficiency. Efficacy means the ability to produce the desired results irrespective of the resources expended. On the other hand, efficiency is not only concerned with achieving results but also concerned with the resources utilised in achieving the results.

This paper details measures that could be used to evaluate efficacy, complexity and efficiency of detection algorithms; as well as their strengths and weaknesses. In addition, it critically analyses the benefits as well as limitations of standardising and customising performance metrics for feedback algorithms in BMI, such that algorithms could be assessed in a fairer, more balanced and objective way.

EFFICACY MEASURES

We review various methods that are currently used in assessing the efficacy of detection algorithms in brain machine interfaces. Evaluating the efficacy of detection algorithms can be tedious, as there are several nearly similar measures with rather subtle differences that can be used. However, some methods are more effective for some BMI applications based on the peculiar features of the said application. While others perform better at revealing certain characteristics of the detection algorithms that may ordinarily be less conspicuous. This review will look at several metrics for assessing efficacy, with the intention of deciding which metrics are ideally tailored for BMI applications. Nevertheless, more emphasis is placed on metrics that are good at evaluating binary decision problems, like disease and non-disease states; actions or inactions; intentions and non-intentions.

A potential pitfall in analysing binary detection algorithms is data redundancy (Mohammed, Zamani, Bayford, & Demosthenous, 2017). That is, using very similar sequence of examples to train and test an algorithms could lead to having an overestimated performance, because it only ends up producing the output to a possible training input rather than interpolating or extrapolating, which is the main aim of a detection algorithm (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). The emphasis here will be on performance evaluation metrics rather than data selection method. At its most basic level, the efficacy of binary detection can be assessed using the confusion matrix. Others are: difference and information measures.

Confusion Matrix

The confusion matrix or contingency table, is a table that contains information on the actual and detected states for a supervised learning classifier. For a binary classifier it is a 2 × 2matrix consisting of the shaded values in Table I. However, when the detection problem is an *M*-ary detection problem, then the confusion matrix can be represented by an M × M matrix. From the confusion matrix in Table I, the various quantities are defined as follows: TP represents the true positives, FP represents the false positives, FN represents the false

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

negatives and TN represents the true negatives. Disease are periods in which the BMI systems detect the on-set of neuro-motor symptoms and non-disease are disease free periods. However, using a range of numbers to represent performance as in the confusion matrix, it may not be obvious how well a detection algorithm performs. An alternative will be to use single performance measures based on a combination of two or more relevant quantities from the confusion matrix. There is no single measure that gives a completely fair assessment of the confusion matrix. Nevertheless, there are measures that give a very good picture such as F1-score and Mathews correlation coefficient (MCC). They give a very balanced measure for uneven classes. Below is a brief description of the various performance metrics obtainable from the confusion matrix that has been used in various literature on BMI.

Table I: Confusion Matrix summarising performance metrics

		Actual State		
		Disease	**Non-Disease**	
Detected State	**Disease**	TP	FP (type I error)	Precision or Positive Predictive Value ($\frac{TP}{TP+FP}$)
	Non-Disease	FN (type II error)	TN	Negative Predictive Value ($\frac{TN}{TN+FN}$)
		Sensitivity or Recall or True Positive Rate ($\frac{TP}{TP+FN}$)	Specificity or True Negative Rate ($\frac{TN}{TN+FP}$)	Classification accuracy, ($\frac{TP+TN}{TP+FN+TN+FP}$) Classification error = 1 - Classification accuracy
		False negative rate or Miss rate ($\frac{FN}{TP+FN}$)	False positive rate or Fall out ($\frac{FP}{TN+FP}$)	

Classification Accuracy and Classification Error

Classification accuracy measures the proportion of correctly classified (detected) states. On the other hand, classification error measures the proportion of classifications that were wrong. Both classification accuracy and error are not suitable metrics for measuring classifier quality in applications with skewed classes. To overcome this, newer metrics like the weighted classification error (WCE), which applies various weights to errors in the different classes depending on which error is more likely to occur and which is more (or less) desirable(Lemm, Blankertz, Dickhaus, & Müller, 2011). WCE can be represented mathematically as,

$$WCE = \left(\frac{\alpha \times FP}{FP + TN}\right) + \left(\frac{(1-\alpha) \times FN}{TP + FN}\right) \qquad (1)$$

where α is the weight (or penalty) applied to error in the disease class and $1-\alpha$ is the weight applied to error in the non-disease class. The error in the disease class represents type I error (false-positive rate), while the error in the non-disease class represents type II error (false-negative rate).

Sensitivity and Specificity

The sensitivity or recall of a detection test gives the percentage of

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

patients for whom the outcome is positive that are correctly detected by the test – true positive rate. On the other hand, the specificity gives the percentage of patients for whom the outcome is negative that are correctly detected by the test – true negative rate.

Positive and Negative Predictive Value

Precision or positive predictive value (PPV) of a test is the probability that a patient has a positive outcome given that they have a positive test result. This is different from sensitivity, which is the probability that a patient has a positive test result given that they have a positive outcome. Similarly, the negative predictive value (NPV) is the probability that a patient has a negative outcome given that they have a negative test result, in contrast to specificity, which is the probability that a patient has a negative test result given that they have a negative outcome.

Miss and Fall-out Rate

The miss rate gives the conditional probability of a negative test result given that the condition being looked for is positive. The fall out rate gives the conditional probability of detection being positive given an event that it was not positive. It gives the significance level of the test. In a sense, the miss and fall-out rate are the opposites of sensitivity and specificity, respectively.

Choice Probability (CP)

ROC is a plot of sensitivity and false positive rate and has an area under the curve (AUC) of between 0 and 1. It is used to evaluate the performance of various detection algorithms. The AUC of the ROC, also called the choice probability (CP), represents the probability that the detector will correctly classify an event in a two-alternative forced-choice classification. This is described in more detail in(Fawcett, 2006).

F1 score

In a model consisting of skewed classes, for example disease= 2% and non-disease= 98%, classification accuracy is not a good metric for analysing such models as a randomly guessing classifier that classifiers all test cases as non-disease will achieve 98% accuracy. This is obviously misleading. In such situation, a single useful metric that can be used to analyse the performance of a classifier is the F1 score, which is obtained using the precision and sensitivity (or recall). Mathematically(Marshland, 2015),

$$F_1 = \frac{2 \times P \times R}{P+R} \qquad (2)$$

From the equation, the F1 score uses the precision and recall to determine the performance of a classifier with a higher weight placed on the lower of the two. This is because; ideally high precision and high recall are required for disease detection in BMI so as to properly modulate therapy. High precision is required because wrong prediction will result in administering stimulation when it is not required, and this could possibly lead to side effects in some neuro-motor disorders (Baizabal-Carvallo & Jankovic, 2016). On the other hand, high recall is also required so as to avoid missing too many disease cases. As a high number of missed cases could worsen the patient state, especially in conditions such as PD, since the patient will be starved of the required stimulation needed to mitigate the effects of disease (Hacker et al., 2015).

Mathews Correlation Coefficient (MCC)

Another balanced measure for disproportionate classes is the MCC. It is a modified version of the Pearson correlation coefficient. It has a range between –1 (total disagreement) and +1 (total agreement). It is normally 0 for totally random predictions, producing a value of 0 for completely independent variables. Mathematically,

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+F\ )(TP+FP)(TN+FP)(TN+FN)}} \quad (3)$$

The major shortcoming of MCC is that it can only be used when one of the denominators *TP + FN, TP + FP, TN + FP and TN + FN* is not a zero(Marshland, 2015). This makes the F1- score a more practical measure, when asses detection algorithms for test cases with uneven classes.

DIFFERENCE AND INFORMATION MEASURES

These are measures that quantify relative discrepancies between the actual results and detected results. Difference measures are performance measures that identify the number of positions (or places) at which predictions differ from the actual results. Weighted distance measures can be used in evaluating skewed classes to account for the disproportionate classes or as a custom accuracy measure in balanced classes. Information measures generally give an insight into the amount of information lost when a predictive model is used to approximate the actual model.

Hamming Distance

This gives the sum of the FP and FN between the actual and predicted results. It compares the two strings of predicted and actual results consisting of 0s (false) and 1s (true), so as to identify the number of positions they differ. Literally, it presents the same result as the classification error, however without being normalised by the total number of test cases. Like the classification error, it is a weak measure when used for test cases with disproportionate classes.

$$HD\ (A, P) = \sum_i |a_i - p_i| \quad (4)$$

Euclidean Distance and L^P D istances

The Euclidean and L^P distances are used to overcome the shortcomings of the Hamming distance. They are mostly used in applications were high level of accuracy may be required. Mathematically,

$$ED\ (A, P) = \sqrt{\sum_i |a_i - p_i|^2} \quad (5)$$

In a similar vein, the L^P distance is defined as,

$$LD\ (A, P) = \sqrt[P]{[\sum_i |a_i - p_i|^P]} \quad (6)$$

The Hamming distance is strictly an L^1 distance, while the Euclidean distance is an L^2 distance.

Projection Error

In dimensionality reduction, data is required to be projected unto a subspace in which majority of the features in the original data are retained. This is measured using the projection error (also called reconstruction error).The formula for obtaining the proportion of variance lost (P_{error}) after projection is,

$$P_{error} = \frac{\sum_{i=1}^{m} \|x^{(i)} - x_{proj}{}^{(i)}\|^2}{\sum_{i=1}^{m} \|x^{(i)}\|^2} \quad (7)$$

Where m is the number of training examples, $\boldsymbol{x} \in \mathbb{R}^m$ are the observations, $\boldsymbol{x_{proj}} \in \mathbb{R}^m$ are the new projections of $\boldsymbol{x} \in \mathbb{R}^m$. The projection error gives the normalised sum of squared error.

Information Gain or Relative Entropy

Relative entropy is a measure of the non-symmetric difference between probability distributions for a true model and a predictive model. As detection accuracy measures they could be used to determine how well the theoretical model (Y) based on detection algorithms approximates the true distribution (X) of data. For an M-ary detection, having a true distribution of probabilities X = (x_1, x_2… x_M) and a theoretical distribution Y = (y_1, y_2… y_M), with probabilities x_i, $y_i \geq 0$ and $\Sigma x_i = \Sigma y_i = 1$; is given as,

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng *Department of Electronics/Electronics Engineering, Nigeria Defence Academy.*

$$HD\ (X,Y) = \sum_i^M x_i \log \frac{x_i}{y_i} \qquad (8)$$

Relative entropy is a useful metric for evaluating how well a classifier discriminates between data instances (Battiti, 1994), (Fisher, Siracusa, & Tieu, 2009). In the strict sense, the relative entropy gives information that is nearly similar to the classification error and Hamming distance. The choice of which one to adopt is dependent on the type of application. For balanced classes, the relative entropy is a performance indicator that is ideally suited for non-binary data, more useful for regression problems rather than classification problems. While the hamming distance and classification error both work well for binary and multiple classes. However, the Hamming distance is suitable in applications with standard strings or sequences like in prediction of proteins secondary structure and secretory signal peptides (Baldi et al., 2000), because the error are not normalised to a standard. Whereas, classification error is more universal because it could be used in any application as long as the total number of test cases are known.

Akaike and Bayesian Information Criterion

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are measures of model selection that selects the most parsimonious model from a finite set of models. Inspired by the information gain from information theory and Fisher's maximised log-likelihood function, both AIC and BIC measures the relative quality of a statistical model from a given set of data (Aho, Derryberry, & Peterson, 2014). It offers a relative estimate of the information lost when a given model is used to represent the 'true model'. When fitting models, the likelihood of a model could be increased by adding parameters. However, doing so may cause overfitting. As a balance between the bias introduced by maximum likelihood and information gain, both the AIC and BIC introduce a penalty term for the number of parameters in the model, with the BIC providing a larger penalty. This is in order to produce a resource efficient model for data analysis. The AIC and BIC can be defined mathematically as,

$$AIC\ =\ -2\ln \mathcal{L}\left(\hat{\theta}\right) + 2p \qquad (9)$$

$$BIC\ =\ -2\ln \mathcal{L}\left(\hat{\theta}\right) + p\ln n \qquad (10)$$

Where $\mathcal{L}\left(\hat{\theta}\right)$ is the likelihood of the estimated model (in the context of general linear models, e.g regression and ANOVA, this is the likelihood of the parameters in a normally distributed model with mean ($\mu = 0$), and variance (σ^2)), p is the total number of parameters that are estimated in the model, and n is the sample size. In both AIC and BIC, smaller values indicate better predictive models.

Both AIC and BIC balance complexity with goodness of fit, however they have some subtle differences. The AIC is better when used to evaluate the best predictive model for a complex processes (Vrieze, 2012). A complex process being a process in which the number of parameters in the true model is d, while the predictive models with parameters in the region of p, where $d >> p$. On the other hand, the BIC does a better job of evaluating the best predictive problem in models where $d << p$. The larger and increasing penalty of BIC, ensures that as the complexity of the process increases (which increases its parameter space), it consistently produces an effective but not necessarily efficient model. Whereas the AIC with a fixed penalty, ensures that the most efficient model is always selected. The choice of AIC or BIC as a measure of efficacy, is a choice between selecting a model that ensures efficiency, which the AIC does, or a model that ensures consistency in efficacy, which the BIC does better. This depends on what the priority is for a given application.

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

Cross-correlation

The cross-correlation of two data sequences measures the similarity of two sequence as a function of the displacement of one relative to the other. It is sometimes called the sliding dot product or sliding inner product. It can be represented mathematically as,

$$Corr_{xy}[k] = \sum_{m=0}^{N-1} x[m]y[m-k] \quad (11)$$

Where k is an integer, $0 \leq k \leq N-1$. $Corr_{xy}[k]$ is the cross correlation at sequence k, $x[k]$ and $y[k]$ are the two data sequences. It is mostly used for searching a long signal for a shorter, like in genomic sequences (Skutkova, Vitek, Sedlar, & Provaznik, 2015). Table II summarises the strengths and weaknesses of different efficacy measures.

Table II: Strengths and weaknesses of various efficacy measures

Efficacy measure		Strength	Weakness	Ref.
Confusion Matrix	Sensitivity, Specificity and Precision	Provides information on only one characteristic. This may come handy in applications where a single efficacy characteristic is relevant.	Seriously misleading when used to assess the overall performance of a detection algorithm.	(Fatourechi, Ward, & Birch, 2008)
	Classification accuracy and error	Simple, straightforward and popular.	Can be very misleading for skewed classes. Weighted classification error can be used to address this.	(Muraskin et al., 2017)
	Choice Probability and ROC	An efficient way to obtain information on the relationship between sensitivity and specificity. A fairly balanced measured for classification with skewed classes.	AUC dependent on curve fitting method. Areas without patient data can be highly extrapolated.	(Rouse et al., 2011)
	F1-score	A balanced measure for classification with skewed classes. Also, provides information on precision and sensitivity.	The lower of the two between sensitivity and precision dominates and influences the measure.	(Yoo, Kim, Filandrianos, Taghados, & Park, 2013)
	MCC	A balanced measure for classification with skewed classes.	It can only be used when one of the denominators *TP + FN, TP + FP, TN + FP and TN + FN* is not a zero.	(Kostoglou et al., 2016)
Difference Measures	Hamming distance	Easy to implement.	Misleading when used to compare scenarios with dissimilar number of detection instances.	(Fukami, Shimada, Forney, & Anderson, 2012)
	L^p distances	The p term could serve as a good penalty term for applications demanding high performance.	The choice of p could overstate or understate performance.	(Goñi et al., 2014)

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

Efficacy measure		Strength	Weakness	Ref.
	Projection error	It measures how well the low dimensional representation captures the features in the original data.	It can be computationally intensive when data consists of a large number of observations as well as dimensions.	(Miyawaki et al., 2008)
Information Measures	Information gain	Captures probabilistic information which is important in applications with a defined probability distribution.	It is a relative measure. So it only accurately measures relative convergence to a specific model.	(Akce, Norton, & Bretl, 2015)
	AIC	Not only measures efficacy, but parsimony as well.	Inadequate for assessing simple processes.	(Schalk, Brunner, Gerhardt, Bischof, &Wolpaw, 2008)
	BIC	Not only measures efficacy, but parsimony as well.	Inadequate for assessing complex processes.	(Yargholi & Hossein-Zadeh, 2016)
	Cross correlation	A simple and straightforward way of comparing data sequences.	Can be misleading when used to compare performance on signals with small and large amplitudes. This exaggerates the large amplitude signal due to the dot product. A possible solution to this is by using the normalised cross correlation.	(Skutkova et al., 2015)

COMPLEXITY MEASURES

The complexity of a BMI detection algorithm is dependent on how computationally intensive the algorithms is. Real time detection in BMI algorithms can only be realised with low computation algorithms. To facilitate implantation this algorithms need to be optimised such that their microchip implementations have significantly small scales in terms of size and power consumption. By leveraging on state of the art CMOS technology and clever computational techniques (low complexity, high speed and low-power); on-line BMI detection processors can be implemented. This could lead to efficient BMI systems that can be deployed for chronic clinical use.

Arithmetic Operations

A critical consideration in processor designs for BMI is to determine the number of operations required to decode information from neural activity. Arithmetic operations are required on recorded signals in order to transform them into useful, manageable or computationally efficient forms (Hebb et al., 2014). Detection algorithms for BMIs mainly involve three stages: feature extraction, dimensionality reduction and pattern classification. Bio-signal processing mainly involves operators that conduct time-domain processing, spectral

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

processing and a mixture of both in some cases. These processing are mainly done using fixed-point or floating-point number representation which are outside the scope of this review. However, arithmetic operations are seen through the lenses of addition and multiplication, which form the basis for all arithmetic operations. The total number of arithmetic operations in digital signal processors is dependent of the resolution of the digital converter used. Minimising complexity without compromising performance can be achieved by optimising the resolution of the digital converter. Also, complexity can be reduced by discarding (or reformulating) all redundant variables (or parameters) in a computation(Lee, Kung, & Verma, 2012). Computationally efficient algorithms are those that scale logarithmically with an increase in a dependent variable.

Hardware Resources

Detection algorithms for BMI systems have essentially carried out off-line processing by tethering patients/participants or wirelessly transferring the data to high performance computers. Tethering is impractical as it hinders normal patient activities. And wirelessly transferring data comes with bandwidth and power restrictions that may not be easily met. Hardware resources are determined by key parameters like power and area. For low-power implementation, there is a shift towards the use of microchip technologies that embed signal processing. Recent advancements in microchip technology have led to the development of detection algorithms for BMI applications. Table III summarises microchip implementations of DBS systems published from 2014 to date. This systems consisted of neural signal processors. All the systems are proof-of-concept implementations using mostly offline processing and were mostly used for closed-loop DBS control of epileptic seizures(Kassiri et al., 2017),(Kassiri et al., 2016),(Biederman et al., 2015), (Shulyzki et al., 2015), (Chen et al., 2014), (Rhew et al., 2014).From Table III, it is clear that applications using signals with low spatial scale (and high temporal resolution) e.g. spikes (unit activity) employ less processing area and power than applications with much large spatial scale (and low temporal resolution) like ECoG and EEG. These confirm the view expressed (Sellers, Krusienski, McFarland, Vaughan, & Wolpaw, 2006), that event-related potentials have shown to be more effective in BMI applications compared to spontaneous signals such as EEG because they do not require large storage requirements and lengthy training periods. The requirements for chronically implanted microchips are becoming more and more stringent to avoid neural tissue damage. This makes the need for minimising power and is in state-of-the-art microprocessors ever more important.

Table III: Closed-loop DBS systems with online recording and stimulation

Ref.	Year	On-chip Stages	Neural Signals	CMOS Tech.	Processing	Power (µW)	Area (mm^2)
(Elhosary et al., 2019)	2019	Recording, feature extraction and classification	EEG	65 nm	ASIC	14 900	0.2022
(Mohammed & Demosthenous, 2018)	2018	feature extraction and detection	LFP	45 nm	FPGA	18.1	1.91

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

Ref.	Year	On-chip Stages	Neural Signals	CMOS Tech.	Processing	Power (µW)	Area (mm^2)
(Kassiri et al., 2017)	2017	Recording, feature extraction and stimulation	EEG/ECoG	0.13 µm	FPGA	1 300	12
(Liu, Zhang, Richardson, Lucas, & Van der Spiegel, 2016)	2017	Feature extraction and stimulation.	Spike and LFP	0.18 µm	CPU	896	3.7
(Kassiri et al., 2016)	2016	Recording and stimulation	ECoG	0.13 µm	FPGA	2 170	16
(Biederman et al., 2015)	2015	Feature extraction, data compression and stimulation.	Spike	65 nm	CPU	626.4	4.8
(Shulyzki et al., 2015)	2015	Recording and stimulation	ECoG	0.35 µm	FPGA	13 500	12.8
(Chen et al., 2014)	2014	Recording, seizure detection and stimulation.	iEEG	0.18 µm	Onsite	2 800	13.5
(Rhew et al., 2014)	2014	Recording, feature extraction and stimulation	Spike and LFP	0.18 µm	Partly onsite	468	4

EFFICIENCY AND OPTIMALITY MEASURES

This introduces efficiency metrics used in evaluating neural signal processors for detection. In detection algorithms, efficiency metrics or measures provide information on the trade-off between resource utilisation and the performance of detection algorithms.

Energy Efficiency

Energy efficiency is a metric defined around operations like multiplications, additions or delays. This is the number of useful operations divided by the energy required to do them. It is measured in operations/Joules and it gives the average energy required for each operation. It is a metric that can be used to assess how optimal an algorithm is compared to other algorithms performing the same function (Marković & Brodersen, 2012).

Energy-Delay Product (EDP)

EDP gives the average energy consumed per switching event. In BMI detection algorithms, switching events are the decisions made by the algorithms. A variation of the EDP is the power-delay

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

product (PDP), which gives the power consumed per switching event. PDP is a misleading metric compared to EDP because it favours a processors that operates at lower frequencies (Rabaey, Chandrakasan, & Nikolic, 2002). EDP gives the average energy multiplied by the time it takes to do the computation. Thus, EDP takes performance into account.

Area Efficiency

With the increase in lab-on-chip techniques, there is an ever increasing need to miniaturise platforms for clinical diagnostics. The need for miniaturisation is becoming inevitable because of the proliferation of chronically implanted devices at very invasive regions in the body. Miniaturisation ensures that the obstruction of normal internal and/or metabolic activity is mitigated (Marković & Brodersen, 2012). Area efficiency is a metric that provides information on the useful area per operation in a microchip. It is obtained by dividing the number of operations by the chip area occupied to do those operations.

TO STANDARDISE OR TO CUSTOMISE PERFORMANCE MEASURES?

A major hurdle in the deployment of standalone BMI systems are the clinical trial costs involved in developing such systems. Due to their sensitive nature, medical devices require excessive regulatory oversight; well thought-through deployment strategy; substantial design and manufacturing cost and most importantly, a rigorous validation process. Generally, the validation process starts prior to preclinical experiments, because for clinical translation to field, it is necessary to ensure that devices perform desired function with high certainty. A major bottleneck in validating algorithms is the choice of performance measures and metrics.

Accurate selection of appropriate performance metrics are the cornerstone for a balanced, objective and fair evaluation of detection algorithms. Performance evaluation is essential in BMI; however, it varies from application to application. Due to the multitude of non-standard performance measures in use for BMI, the performance of the same algorithm can be interpreted differently. For example, a single algorithm can have a 'high' measure in one metric and a 'low' measure in another metric, both of which intend to measure the same property. Inconsistencies like this further obfuscate the validation process. This is not far-fetched considering the rise in use of words like 'high/increased accuracy', 'low/reduced delay', 'low/reduced complexity' and so on to describe BMI performance. Words like these, instead of providing more insights, they end up creating more unanswered questions like; what sort of efficacy or complexity metrics were adopted? How do the selected metrics compare to other adopted metrics? How beneficial are these measures/metrics to the application? How does this system compare to other current systems that use a different metric? All these questions could be avoided if standardised measures are used, such that every current work adopts a standard metric. Another common issue that reinforces the argument for standardisation, is that performance measures with the same meaning sometimes have different names. A typical example of this is the classification accuracy sometimes called success rate, mean accuracy, non-error rate, probability of success and many more names. This inconsistency in nomenclature can lead to misinterpreting the metric and discrepancies in reporting performance.

To counter the argument for having standardised measures for BMI applications, customised measures can be created for various applications. Customised measures using weighted efficacy and complexity measures could be used to more accurately measure

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

application specific efficiency or optimality. This was done in (Mohammed et al., 2017), were complexity and efficacy were assessed using customised measures; and the algorithms with the best trade-offs in terms of efficacy and complexity were selected for implementation. This is necessary since the efficacy needs of various applications are dissimilar as highlighted and issues like complexity affect applications differently depending on the need for chronic or non-chronic implantation; and how invasive, semi-invasive or non-invasive the implant is. For customising performance metrics, metrics that emphasise more on the current needs and trends of the industry can be selected. Currently, efficacy is of primary importance as most studies are proof-of-concept studies.

Until BMIs achieve performance levels that allow patients accomplish routine activities unassisted, efficacy still remains the most important requirement for BMI systems. Afterwards, focus can then be shifted to other metrics that asses complexity and efficiency.In spite of the major breakthroughs in BMI, there is still a long way to go before they can find large scale application(Pisotta, Perruchoud, & Ionta, 2015). As stated earlier, their applicability is highly dependent on ease of use. Similarly, the adoption of different metrics for evaluating detection algorithms in BMI is dependent on their ease of measurement and how universal they are. It is thus recommended that performance measures should be selected based on the research issues addressed by each individual application.

CONCLUSION

For BMI feedback algorithms, a major challenge is the unavailability of standard measures for evaluating feedback algorithms in BMI applications. For optimal performance, detection algorithms in BMI systems are required to have high efficacy, low complexity and high efficiency. To ascertain these requirements, the first challenge is to choose which metrics or measures to use as performance indicators. So far in BMI applications, performance requirements are to an extent dependent on the type of application. Because, a measure that is evidently 'high' in one application may be evidently 'low' in some other application. Thus, standardising and/or customising metrics for evaluating performance still remains a big challenge. Nevertheless, various applications have gravitated towards certain metrics and standards based on their simplicity and popularity of use in state of the art designs. An example of such is the use of classification accuracy. For all its inconsistencies, the classification accuracy seems to be the most widely used metric for measuring efficacy. This can be attributed to its simple and straightforward nature – even though it is sometimes misleading. However, for a more accurate assessment of efficacy, more balanced measures like F1-score, MCC and many more customised metrics that incorporate one or more reliable efficacy measures could be used.

Finally, for BMIs to facilitate unrestricted interaction by patients, the goal still remains real-time detection algorithms with hardware efficiency at the highest achievable accuracy. However, current research is still miles away from this mark. The major bottleneck is the great scientific challenge of understanding how the brain works. Notwithstanding, there seems to be so much potential in BMI applications. Who knows, BMI applications could take a dramatic turn in fortunes such that 100% performance levels are achieved in our lifetime.

REFERENCES

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636.

Akce, A., Norton, J. J. S., & Bretl, T. (2015).

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

An SSVEP-Based Brain-Computer Interface for Text Spelling With Adaptive Queries That Maximize Information Gain Rates. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(5), 857–866. https://doi.org/10.1109/TNSRE.2014.2373338

Baizabal-Carvallo, J. F., & Jankovic, J. (2016). Movement disorders induced by deep brain stimulation. *Parkinsonism & Related Disorders*. https://doi.org/10.1016/j.parkreldis.2016.01.014

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics (Oxford, England)*, *16*(5), 412–424.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550. https://doi.org/10.1109/72.298224

Biederman, W., Yeager, D. J., Narevsky, N., Leverett, J., Neely, R., Carmena, J. M., … Rabaey, J. M. (2015). A 4.78 mm 2 Fully-Integrated Neuromodulation SoC Combining 64 Acquisition Channels With Digital Compression and Simultaneous Dual Stimulation. *IEEE Journal of Solid-State Circuits*, *50*(4), 1038–1047. https://doi.org/10.1109/JSSC.2014.2384736

Chen, W.-M., Chiueh, H., Chen, T.-J., Ho, C.-L., Jeng, C., Ker, M.-D., … Wu, C.-Y. (2014). A Fully Integrated 8-Channel Closed-Loop Neural-Prosthetic CMOS SoC for Real-Time Epileptic Seizure Control. *IEEE Journal of Solid-State Circuits*, *49*(1), 232–247. https://doi.org/10.1109/JSSC.2013.2284346

Elhosary, H., Zakhari, M. H., Elgammal, M. A., Abd El Ghany, M. A., Salama, K. N., & Mostafa, H. (2019). Low-Power Hardware Implementation of a Support Vector Machine Training and Classification for Neural Seizure Detection. *IEEE Transactions on Biomedical Circuits and Systems*, *13*(6), 1324–1337. https://doi.org/10.1109/TBCAS.2019.2947044

Fatourechi, M., Ward, R. K., & Birch, G. E. (2008). A self-paced brain–computer interface system with a low false positive rate. *Journal of Neural Engineering*, *5*(1), 9–23. https://doi.org/10.1088/1741-2560/5/1/002

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fisher, J. W., Siracusa, M., & Tieu, K. (2009). Estimation of Signal Information Content for Classification. *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, 353–358. https://doi.org/10.1109/DSP.2009.4785948

French, D. D., Campbell, R. R., Sabharwal, S., Nelson, A. L., Palacios, P. A., & Gavin-Dreschnack, D. (2007). Health care costs for patients with chronic spinal cord injury in the Veterans Health Administration. *The Journal of Spinal Cord Medicine*, *30*(5), 477–481. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18092564

Fukami, T., Shimada, T., Forney, E., & Anderson, C. W. (2012). EEG character identification using stimulus sequences designed to maximize mimimal hamming distance. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1782–1785. https://doi.org/10.1109/EMBC.201

Corresponding author: Ashraf, A. A. aaashraf@nda.edu.ng Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

2.6346295

Goñi, J., van den Heuvel, M. P., Avena-Koenigsberger, A., Velez de Mendizabal, N., Betzel, R. F., Griffa, A., … Sporns, O. (2014). Resting-brain functional connectivity predicted by analytic measures of network communication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(2), 833–838. https://doi.org/10.1073/pnas.1315529111

Hacker, M. L., Tonascia, J., Turchan, M., Currie, A., Heusinkveld, L., Konrad, P. E., … Charles, D. (2015). Deep brain stimulation may reduce the relative risk of clinically important worsening in early stage Parkinson's disease. *Parkinsonism & Related Disorders*, *21*(10), 1177–1183. https://doi.org/10.1016/j.parkreldis.2015.08.008

Hebb, A. O., Zhang, J. J., Mahoor, M. H., Tsiokos, C., Matlack, C., Chizeck, H. J., & Pouratian, N. (2014). Creating the Feedback Loop: closed-loop neurostimulation. *Neurosurgery Clinics of North America*, *25*(1), 187–204. https://doi.org/10.1016/j.nec.2013.08.006

Kassiri, H., Bagheri, A., Soltani, N., Abdelhalim, K., Jafari, H. M., Salam, M. T., … Genov, R. (2016). Battery-less Tri-band-Radio Neuro-monitor and Responsive Neurostimulator for Diagnostics and Treatment of Neurological Disorders. *IEEE Journal of Solid-State Circuits*, *51*(5), 1274–1289. https://doi.org/10.1109/JSSC.2016.2528999

Kassiri, H., Tonekaboni, S., Salam, M. T., Soltani, N., Abdelhalim, K., Velazquez, J. L. P., & Genov, R. (2017). Closed-Loop Neurostimulators: A Survey and A Seizure-Predicting Design Example for Intractable Epilepsy Treatment. *IEEE Transactions on Biomedical Circuits and Systems*, 1–15. https://doi.org/10.1109/TBCAS.2017.2694638

Kostoglou, K., Michmizos, K. P., Stathis, P., Sakas, D., Nikita, K. S., & Mitsis, G. D. (2016). Classification and Prediction of Clinical Improvement in Deep Brain Stimulation from Intraoperative Microelectrode Recordings. *IEEE Transactions on Biomedical Engineering*, 1–1. https://doi.org/10.1109/TBME.2016.2591827

Lee, K. H., Kung, S.-Y., & Verma, N. (2012). Low-energy Formulations of Support Vector Machine Kernel Functions for Biomedical Sensor Applications. *Journal of Signal Processing Systems*, *69*(3), 339–349. https://doi.org/10.1007/s11265-012-0672-8

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Liu, X., Zhang, M., Richardson, A. G., Lucas, T. H., & Van der Spiegel, J. (2016). Design of a Closed-Loop, Bidirectional Brain Machine Interface System With Energy Efficient Neural Feature Extraction and PID Control. *IEEE Transactions on Biomedical Circuits and Systems*, 1–14. https://doi.org/10.1109/TBCAS.2016.2622738

Marceglia, S., Rossi, E., Rosa, M., Cogiamanian, F., Rossi, L., Bertolasi, L., … Priori, A. (2015). Web-based telemonitoring and delivery of caregiver support for patients with Parkinson disease after deep brain stimulation: protocol. *JMIR Research Protocols*, *4*(1), e30. https://doi.org/10.2196/resprot.4044

Marković, D., & Brodersen, R. (2012). *DSP Architecture Design Essentials*. New York, NY: Springer.

Corresponding author: Ashraf, A. A. ✉ aaashraf@nda.edu.ng ✉ Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

Marshland, S. (2015). *Machine Learning : an algorithmic perspective* (2nd ed.). Boca Raton, FL: CRC Press.

Mason, S. G., & Birch, G. E. (2003). A general framework for brain-computer interface design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *11*(1), 70–85. https://doi.org/10.1109/TNSRE.2003.810426

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron*, *60*(5), 915–929. https://doi.org/10.1016/j.neuron.2008.11.004

Mohammed, A., & Demosthenous, A. (2018). Complementary Detection for Hardware Efficient On-Site Monitoring of Parkinsonian Progress. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *8*(3), 603–615. https://doi.org/10.1109/JETCAS.2018.2830971

Mohammed, A., Zamani, M., Bayford, R., & Demosthenous, A. (2017). Toward On-Demand Deep Brain Stimulation Using Online Parkinson's Disease Prediction Driven by Dynamic Detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(12), 2441–2452. https://doi.org/10.1109/TNSRE.2017.2722986

Muraskin, J., Sherwin, J., Lieberman, G., Garcia, J. O., Verstynen, T., Vettel, J. M., & Sajda, P. (2017). Fusing Multiple Neuroimaging Modalities to Assess Group Differences in Perception–Action Coupling. *Proceedings of the IEEE*, *105*(1), 83–100. https://doi.org/10.1109/JPROC.2016.2574702

Pisotta, I., Perruchoud, D., & Ionta, S. (2015). Hand-in-hand advances in biomedical engineering and sensorimotor restoration. *Journal of Neuroscience Methods*, *246*, 22–29. https://doi.org/10.1016/j.jneumeth.2015.03.003

Rabaey, J. M., Chandrakasan, A. P., & Nikolic, B. (2002). *Digital Integrated Circuits - A Design Perspective* (2nd ed.). Pearson.

Rhew, H., Jeong, J., Fredenburg, J. A., Member, S., Dodani, S., Patil, P. G., … Member, S. (2014). *A Fully Self-Contained Logarithmic Closed-Loop Deep Brain Stimulation SoC With Wireless Telemetry and Wireless Power Management*. 1–15.

Rouse, A. G., Stanslaski, S. R., Cong, P., Jensen, R. M., Afshar, P., Ullestad, D., … Denison, T. J. (2011). A chronic generalized bi-directional brain–machine interface. *Journal of Neural Engineering*, *8*(3), 036018. https://doi.org/10.1088/1741-2560/8/3/036018

Schalk, G., Brunner, P., Gerhardt, L. A., Bischof, H., & Wolpaw, J. R. (2008). Brain–computer interfaces (BCIs): Detection instead of classification. *Journal of Neuroscience Methods*, *167*(1), 51–62. https://doi.org/10.1016/j.jneumeth.2007.08.010

Secretary, W. H. P. (2013). Fact sheet: BRAIN Initiative. Retrieved February 1, 2017, from https://obamawhitehouse.archives.gov/the-press-office/2013/04/02/fact-sheet-brain-initiative

Sellers, E. W., Krusienski, D. J., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2006). A P300 event-related potential brain–computer interface (BCI): The effects of matrix size and inter stimulus interval on performance. *Biological Psychology*, *73*(3), 242–252. https://doi.org/10.1016/j.biopsycho.2006.04.007

Shanechi, M. M. (2019). Brain–machine interfaces from motor to mood.

Corresponding author: Ashraf, A. A. ✉ aaashraf@nda.edu.ng ✉ Department of Electronics/Electronics Engineering, Nigeria Defence Academy.

Nature Neuroscience, *22*(10), 1554–1564. https://doi.org/10.1038/s41593-019-0488-y

Shulyzki, R., Abdelhalim, K., Bagheri, A., Salam, M. T., Florez, C. M., Velazquez, J. L. P., … Genov, R. (2015). 320-Channel Active Probe for High-Resolution Neuromonitoring and Responsive Neurostimulation. *IEEE Transactions on Biomedical Circuits and Systems*, *9*(1), 34–49. https://doi.org/10.1109/TBCAS.2014.2312552

Skutkova, H., Vitek, M., Sedlar, K., & Provaznik, I. (2015). Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of Theoretical Biology*, *385*, 20–30. https://doi.org/10.1016/j.jtbi.2015.08.007

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Yargholi, E., & Hossein-Zadeh, G.-A. (2016). Brain Decoding-Classification of Hand Written Digits from fMRI Data Employing Bayesian Networks. *Frontiers in Human Neuroscience*, *10*, 351. https://doi.org/10.3389/fnhum.2016.00351

Yoo, S.-S., Kim, H., Filandrianos, E., Taghados, S. J., & Park, S. (2013). Non-Invasive Brain-to-Brain Interface (BBI): Establishing Functional Links between Two Brains. *PLoS ONE*, *8*(4), e60410. https://doi.org/10.1371/journal.pone.0060410

Corresponding author: Ashraf, A. A. ✉ aaashraf@nda.edu.ng ✉ Department of Electronics/Electronics Engineering, Nigeria Defence Academy.