Offline Accuracy: A Potentially Misleading Metric in Myoelectric Pattern Recognition for Prosthetic Control

Max Ortiz-Catalan, *IEEE Member*, Faezeh Rouhani, Rickard Brånemark, and Bo Håkansson

***Abstract*— Offline *accuracy* has been the preferred performance measure in myoelectric pattern recognition (MPR) for the prediction of motion volition. In this study, different metrics relating the fundamental binary prediction outcomes were analyzed. Our results indicate that *global accuracy* is biased by 1) the unbalanced number of possible true positive and negative outcomes, and 2) the almost perfect *specificity* and *negative predicted value*, which were consistently found across algorithms, topologies, and movements (individual and simultaneous). Therefore, *class-specific accuracy* is advisable instead. Additionally, we propose the use of *precision* (*positive predictive value*) and *sensitivity* (*recall*) as a complement to accuracy to better describe the discrimination capabilities of MPR algorithms, as these consider the effect of false predictions. However, all the studied offline metrics failed to predict real-time decoding, and therefore real-time testing continue to be necessary to truly evaluate the clinical usability of MPR.**

I. Introduction

Powered prosthetic devices currently offer more degrees of freedom (DoF) and those possibly controlled in an intuitive manner by patients with missing limbs. The conventional control strategy for such prostheses is based on myoelectric signals recorded by superficial electrodes. The strength of the myoelectric activity captured by one electrode is linked to the actuation of a prosthetic unit in one direction (e.g., *hand open*), while the opposite direction (e.g., *hand close*) is normally controlled by a second electrode, which is placed in a muscle antagonistic to the first one. Because isolation of independent control signals is difficult to achieve by surface electrodes, this *direct control* strategy is normally limited to a single DoF.

Myoelectric pattern recognition (MPR) has been explored as an alternative to *direct control* for decades [1], [2]. It has been shown that a variety of movements, or postures, can be predicted using a set of superficial electrodes, and thus potentially provide intuitive control of several DoF.

The most common metric to evaluate the performance of a given MPR algorithm is the classification accuracy, or its complement, the classification error [3]. However, accuracy can be an ambiguous term which authors tend to obviate and skip its definition when reporting. Since accuracy can be calculated in a variety of ways, with no standardized reporting form [4], misinterpretation can occur and it has been reported problematic in clinical tests [5].

Additionally, classification accuracy computed with pre-recorded data (offline) has been observed to provide a higher expectation of real-time performance than actually delivered [6]–[9]. On the other hand, perfect offline accuracy has been found not always necessary to yield controllable systems [9], [10]. These findings suggest the need for offline MPR performance metrics that can better relate to real-time controllability, such as the *active error rate* (1–*precision*), which has been proposed [11] and further reported [8] as a more informative indicator of MPR performance.

In this work, fundamental metrics for binary classification were evaluated in MPR of individual and simultaneous hand and wrist movements. A variety of MPR algorithms were employed in single and dedicated topologies in order to study the different metrics, as well as their potential relationship to real-time performance.

II. Methods

A. Definitions

For every binary prediction made by the classifier, there are four possible outcomes for each of the classes involved: *true positive* (correct activation), *true negative* (correct inactivation), *false positive* (incorrect activation), and *false negative* (incorrect inactivation). These possible outcomes and their basic ratios are shown in Table I.

		Reality or Condition		
		True	**False**	**Basic Ratio**
Prediction	**Positive**	**T**rue **P**ositive (TP)	**F**alse **P**ositive (FP)	PPV (Precision) $= \frac{TPs}{TPs + FPs}$
	Negative	**F**alse **N**egative (FN)	**T**rue **N**egative (TN)	NPV $= \frac{TNs}{TNs + FNs}$
	Basic Ratio	Sensitivity $= \frac{TPs}{TPs + FNs}$	Specificity $= \frac{TNs}{TNs + FPs}$	

Table I. Possible outcome of binary classification and their basic ratios. PPV = positive predicted value (also known as precision), and NPV = negative predicted value.

MPR is of most value in multi-class problems. This means that there are several movements that can be binary predicted concurrently. A prediction can be considered absolutely correct, if the outcome of all the classes involved is true (positive and negatives), i.e., it is enough for one class to be classified erroneously, for the absolute prediction to be

This work was funded by VINNOVA, CONACYT, and Integrum AB.

M. Ortiz-Catalan is with the Dept. of Signals and System, Chalmers University of Technology (CTH), the Centre for Advance Reconstruction of Extremities (C.A.R.E.), Sahlgrenska University Hospital (SUH), and Integrum AB, all in Gothenburg, Sweden (e-mail: maxo@chalmers.se).

R. Brånemark is with C.A.R.E., SUH, Dept. of Orthopaedics, Gothenburg University (e-mail: rickard.branemark@orthop.gu.se).

F. Rouhani and B. Håkansson are with CTH (e-mail: faezehrouhani@gmail.com and boh@chalmers.se, respectively).

incorrect. For example, consider an algorithm with the task to predict the activation of a myoelectric hand (open/close) and elbow (flexion/extension). If the subjects aims for *hand open*, and the classifier outputs *hand open* and *elbow flexion*, the class *hand open* would be a *true positive* (correct), but the absolute prediction will be incorrect because *elbow flexion* will be a *false positive*.

The latter exemplifies two different ways in which accuracy can be computed.

- **Global accuracy.** This is the most general computation including all possible outcomes for each class:

$$AccG = \frac{TPs + TNs}{TPs + TNs + FNs + FPs} \quad (1)$$

- **Class-specific accuracy.** Alternatively to consider the outcome of each class individually, absolutely correct predictions can be used instead (no false outcomes for any movement) :

$$AccCS = \frac{absolute\ \ correct\ \ prediction\ \ s}{total\ \ absolute\ \ prediction\ \ s} \quad (2)$$

It is a common practice to report the accuracy for each class separately, in which case, attention must be paid to use the proportional amount of the total predictions, *i.e.*, if a total of 120 predictions were made for 6 classes (20 each), the class specific accuracy must considered the outcome of the 20 predictions where that specific class was expected, divided by 20.

Additionally, the outcome of the binary predictions can be combined in basic ratios that provide relevant information on the test capability to identify the true condition (reality) [12], such as:

- **Sensitivity (recall),** which in the context of MPR relates the portion of desired movements (positives) that are predicted as such (true). In other words, if the subject is intending a given motion, how likely is for the classifier to activate it?
- **Specificity,** on the other hand, addresses the question of how likely is for a given movement to stay or go inactive (negative), if not required (true).

Conversely, the predictive values provide information on the correctness of the prediction outcome, rather than the capabilities of the test to identify reality [13].

- **Positive predictive value (PPV) or precision** relates the portion of positive predictions that are true. In MPR, it addresses the question of how likely is for a predicted movement to be the correct one.
- **Negative predictive value (NPV)** addresses the question: If a motion is not predicted (negative), how likely is that it was actually desired?

Additionally, the equilibrium on how *false positives* and *negatives* relate to *true positives* is commonly measured by the F_1-score defined as:

$$F_1 = 2 \times \frac{precision\ \ \times sensitivit\ \ y}{precision\ \ + sensitivit\ \ y} \quad (4)$$

B. Myoelectric pattern recognition

The aforementioned metrics were used to evaluate the prediction of individual and simultaneous hand and wrist movements. Table II summarizes the data used in this study.

Set	Electrodes	Mov.	Sub.
Individual	4	11	20
Simultaneous	8	27	17

Table II. Summary of the myoelectric recordings from individual and simultaneous movements (Mov.) recorded in the most proximal third of the forearm with 4-8 bipolar electrodes equally spaced in 17-20 subjects (Sub.).

The individual movements were: hand open/close, wrist flexion/extension, pro/supination, side grip, fine grip, thumb up, index extension, and *no motion* (11 classes) [9]. The simultaneous movements were: hand open/close, wrist flexion/extension and pro/supination, plus all their possible combinations, and *no motion* (27 classes) [14]. These data sets are freely available online within BioPatRec [9].

The data acquisition, recording protocol, and signal processing are described elsewhere for individual [9] and simultaneous movements [14]. Briefly, four time-domain features (mean absolute value, wave length, zero crossings, and slope sign changes) were extracted from 121 time windows of 200 ms (isometric contractions at comfortable and sustained force level). The resulting feature vectors were divided into 48 (40%) and 24 (20%), for training and validation, respectively. The remaining 49 (40%) feature vectors were used for computing the aforementioned classification metrics. The results here reported corresponded to the average of 10 evaluations for which all the 121 feature vectors were randomized into the different sets (training, validation, and testing) prior to the computation of the evaluation metrics (cross-validation).

Fundamentally different classifiers were employed in this study: Linear Discriminant Analysis (LDA) as a statistical classifier [15]; Multi-Layer Perceptron (MLP) as a supervised Artificial Neural Network (ANN) [16]; Self-Organized Map (SOM) as an unsupervised ANN [16]; and Regulatory Feedback Networks (RFN), a negative-feedback based algorithm [9]. Additionally, the conventional use of these classifiers in a single topology was compared to better performing one-vs-one (OVO) [8], [14], and ago-antagonistic-mixed (AAM) [14] topologies for individual and simultaneous predictions, respectively.

III. Results

Table III summarizes the results for all metrics, and these are graphically presented in box plots where a central mark indicates the median value; the edges are the 25th and 75th percentiles; the whiskers give the range of data values; and solid markers represent the mean values. The standard deviation is not shown in the summary tables for clarity,

however, the results distribution can be observed in Figs 1-4. RFN was omitted from the figures on simultaneous movements due to its poor performance, and LDA is inherently unable to predict simultaneous motions as a single classifier without the *label power set* transformation [14].

(%) AccCS / AccG	Ind. (Single)	Ind. (OVO)	Sim. (Single)	Sim. (AAM)
LDA	92.1 / 98.6	95.9 / 99.2	-	79.0 / 96.4
MLP	90.1 / 98.7	92.8 / 98.7	93.3 / 98.9	94.0 / 98.9
SOM	88.5 / 98.3	94.4 / 99.0	93.7 / 98.8	93.6 / 98.8
RFN	84.0 / 97.1	87.3 / 97.7	17.7 / 81.3	33.8 / 85.4
Sens. / Spec.				
LDA	92.1 / 99.2	95.9 / 99.6	-	93.4 / 98.0
MLP	91.3 / 99.4	92.8 / 99.3	98.0 / 99.4	97.9 / 99.4
SOM	93.2 / 98.8	94.4 / 99.4	98.7 / 98.9	98.5 / 99.0
RFN	84.0 / 98.4	87.3 / 98.7	47.2 / 97.5	76.4 / 89.6
PPV (Prec.) / NPV				
LDA	92.1 / 99.2	95.9 / 99.6	-	93.7 / 96.9
MLP	93.9 / 99.1	92.8 / 99.3	97.9 / 99.0	98.0 / 99.0
SOM	89.1 / 99.3	94.4 / 99.4	96.9 / 99.3	97.1 / 99.3
RFN	84.0 / 98.4	87.3 / 98.7	86.3 / 80.4	72.8 / 90.3
F_1-score				
LDA	92.1	95.9	-	93.5
MLP	92.5	92.8	97.9	97.9
SOM	91.0	94.4	97.7	97.8
RFN	84.0	87.3	60.5	74.6

Table III. Evaluation metrics for the prediction of individual (Ind.) and simultaneous (Sim.) movements.

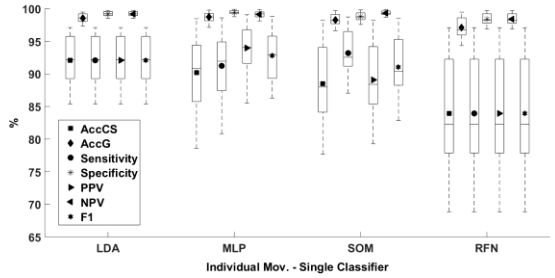

Figure 1. Evaluation metrics for the predicition of 11 individual movements (20 subjects) in a single classifier topology.

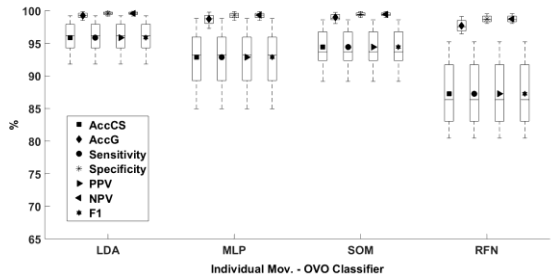

Figure 2. Evaluation metrics for the predicition of 11 individual movements (20 subjects) in a One-Vs-One (OVO) classifier topology.

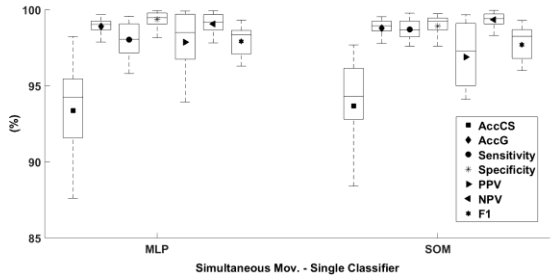

Figure 3. Evaluation metrics for the predicition of simultaneous movements in 3 DoF (27 classes - 17 subject) in a single classifier topology.

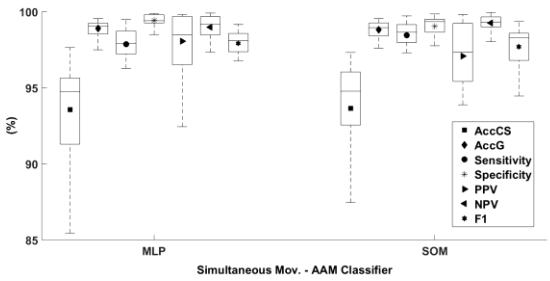

Figure 4. Evaluation metrics for the predicition of simultaneous movements in 3 DoF (27 classes - 17 subject) in a agoantagonistic and mixed (AAM) topology.

In order to investigate how the offline metrics relate to real-time classification, these were compared with the real-time accuracy from the *motion test* reported in [9]. Individual movements predicted by LDA, MLP and RFN, matched with its corresponding subjects and movements are plotted versus real-time accuracy in Fig. 5.

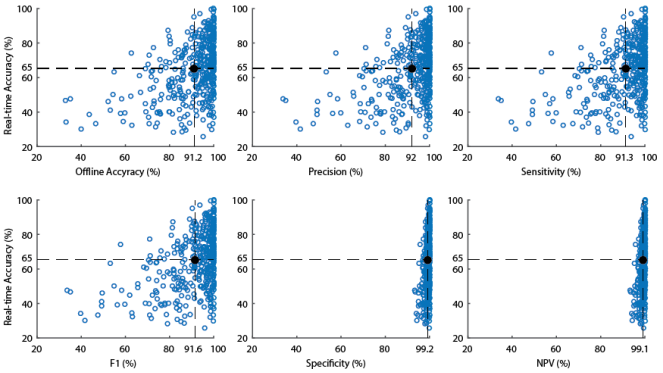

Figure 5. Offline evaluation metrics for individual movements plotted agains real-time accuracy. Predictions by LDA, MLP, and RFN as reported in [9].

IV. Discussion

Our results indicate that a higher *specificity* over *sensitivity* is a common situation in MPR, across classifiers, topologies, and type of movements. Similarly, the *NPV* was consistently found higher than the *PPV* (*precision*). In this particular set of individual movements, for every *true positive,* there are ten *true negatives*, thus the effect of *false* predictions is higher when considering ratios involving *true positives* (*sensitivity* and *PPV*) over *true negatives* (*specificity* and *NPV*). This might explain why the persistent difference between them, which is less accentuated in the prediction of simultaneous movements as the number of possible *true positives* increases.

The almost perfect *specificity* and *NPV* suggests that when a classifier predicts a movement as inactivate (negative), it is very likely that the particular movement was not required. Moreover, no considerably different effect was observed from *false positives* or *negatives* when considering the total number of *true negatives*.

The higher number of possible *true negatives* over *positives*, together with the almost perfect *specificity* and *NPV* showed by all classifiers and topologies, explain why *global accuracy* is almost perfect, and thus potentially misleading. Conversely, *class-specific accuracy* is not affected by the imbalance between the possible *true*

negatives and *positives*, thus it should be preferred over global accuracy. Moreover, attention must be paid as these two computations can produce conflicting results, as in the case of LDA and MLP (Ind. - Single).

In our previous work with MPR [9], [14], [17], [18], we arbitrarily decided to report our findings using class-specific accuracy without defining it, while only few authors have explicitly done so [19], [20]. Regardless of the computation use to calculate accuracy, it is advisable to always define it.

Considering the bias towards higher *global accuracy*, and that *false positives* have been suggested more detrimental to prosthetic control than *false negatives* [11], *precision* and *sensitivity* are potentially more interesting metrics over global accuracy as they consider the effect of false predictions over the less represented true condition (positives). The balance between these metrics can be monitored with the F_1-score, however if given the choice, a higher *precision* over *sensitivity* is potentially preferred to improve controllability. Further real-time evaluations will be conducted to test this hypothesis.

Our results further suggest that one reason for the OVO to outperform single classifiers [8], [14], is mainly due to the improvement of *precision,* which also balances its relationship with *sensitivity* without compromising it (increased F_1-score). Interestingly, LDA and RFN showed on average the same number of *false positive* and *negatives* as single classifiers for individual movements, a behavior observed across all other classifiers in the OVO topology.

V. Conclusion

The conclusions of this work can be summarized as follow:

- Regardless of the computation of accuracy employed, this must always be defined in order to avoid ambiguity.
- High *specificity* and *NPV* are common in MPR for the prediction of motion intent across classifiers, topologies, and types of movements (individual and simultaneous).
- Since *global accuracy* is favored by the latter and the unbalanced true condition, *class-specific accuracy* is preferable instead.
- ***Precision*** (or *active error rate* [11]) can be considered as a critical parameter due to the potential detrimental effect of *false positives* in controllability. Therefore it must be considered along *sensitivity* and *accuracy* in order to provide a more complete report of the discrimination capabilities of a given MPR algorithm.

References

[1] F. Finley and R. Wirta, "Myocoder-computer study of electromyographic patterns," *Arch Phys Med Rehabil*, vol. 48, no. 1, pp. 20–4, 1967.

[2] D. Farina and O. Aszmann, "Bionic Limbs: Clinical Reality and Academic Promises.," *Sci. Transl. Med.*, vol. 6, no. 257, p. 257ps12, Oct. 2014.

[3] E. J. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J Rehabil Res Dev*, vol. 48, no. 6, p. 643, 2011.

[4] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinforma. Rev.*, vol. 16, no. 5, pp. 412–424, 2000.

[5] H. Honest and K. Khan, "Reporting of measures of accuracy in systematic reviews of diagnostic literature," *BMC Health Serv. Res.*, vol. 2, no. 4, 2002.

[6] B. A. Lock, K. Englehart, and B. Hudgins, "Real-time myoelectric control in a virtual environment to relate usability vs. accuracy," in *Proc. MyoElectric Controls/Powered Prosthetics Symposium*, 2005.

[7] G. Li, A. E. Schultz, and T. Kuiken, "Quantifying pattern recognition-based myoelectric control of multifunctional transradial prostheses.," *IEEE Trans Neural Syst Rehabil Eng*, vol. 18, no. 2, pp. 185–92, Apr. 2010.

[8] E. J. Scheme, K. B. Englehart, and B. S. Hudgins, "Selective classification for improved robustness of myoelectric control under nonideal conditions.," *IEEE Trans Biomed Eng*, vol. 58, no. 6, pp. 1698–705, Jun. 2011.

[9] M. Ortiz-Catalan, R. Brånemark, and B. Håkansson, "BioPatRec: A modular research platform for the control of artificial limbs based on pattern recognition algorithms," *Source Code Biol Med*, vol. 8, no. 11, 2013.

[10] N. Jiang, I. Vujaklija, H. Rehbaum, B. Graimann, and D. Farina, "Is Accurate Mapping of EMG Signals on Kinematics Needed for Precise Online Myoelectric Control?," *IEEE Trans Neural Syst Rehabil Eng*, Oct. 2013.

[11] L. J. Hargrove, E. J. Scheme, K. B. Englehart, and B. S. Hudgins, "Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis.," *IEEE Trans Neural Syst Rehabil Eng*, vol. 18, no. 1, pp. 49–57, Feb. 2010.

[12] A. G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity," *Contin Educ Anaesth Crit Care Pain*, vol. 8, no. 6, pp. 221–223, Dec. 2008.

[13] A. K. Akobeng, "Understanding diagnostic tests 1: sensitivity, specificity and predictive values.," *Acta Paediatr.*, vol. 96, pp. 338–41, Mar. 2006.

[14] M. Ortiz-Catalan, B. Håkansson, and R. Brånemark, "Real-Time and Simultaneous Control of Artificial Limbs Based on Pattern Recognition Algorithms," *IEEE Trans Neural Syst Rehabil Eng*, vol. 22, no. 4, pp. 756–764, 2014.

[15] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[16] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River: Prentice Hall, 1999.

[17] M. Ortiz-Catalan, R. Brånemark, and B. Håkansson, "Evaluation of classifier topologies for the real-time classification of simultaneous limb motions," in *35th Ann Int Conf IEEE EMBS*, 2013.

[18] M. Ortiz-Catalan, B. Håkansson, and R. Brånemark, "Real-time classification of simultaneous hand and wrist motions using Artificial Neural Networks with variable threshold outputs," in *Proceedings of the XXXIV International Conference on Artificial Neural Networks (ICANN)*, 2013, pp. 1159–1164.

[19] H. Huang, P. Zhou, G. Li, and T. Kuiken, "Spatial filtering improves EMG classification accuracy following targeted muscle reinnervation.," *Ann. Biomed. Eng.*, vol. 37, no. 9, pp. 1849–57, Sep. 2009.

[20] G. Li, Y. Li, Z. Zhang, Y. Geng, and R. Zhou, "Selection of sampling rate for EMG pattern recognition based prosthesis control.," in *32nd Ann Int Conf IEEE EMBS*, 2010, vol. 2010, pp. 5058–61.