Assessment of EMG Benchmark Data for Gesture Recognition Using the NinaPro Database

Jason Chang, *Student Member, IEEE*, Angkoon Phinyomark, *Member, IEEE*, and Erik Scheme, *Senior Member, IEEE*

***Abstract*— In recent years, many electromyography (EMG) benchmark databases have been made publicly available to the myoelectric control research community. Many small laboratories that lack the instrumentation, access, and experience needed to collect quality EMG data have used these benchmark datasets to explore and propose new signal processing and pattern recognition algorithms. It is widely accepted that noise contamination can affect the performance of myoelectric control systems, and so useful datasets should maintain good signal quality to ensure accurate results for proposed EMG-based gesture recognition systems. Despite the availability and adoption of benchmarks datasets, however, the quality of the EMG signals in these benchmarks has not yet been examined. In this study, the signal quality of the Non-Invasive Adaptive Prosthetics (NinaPro) dataset, the most widely known publicly available benchmark database to date, was comprehensively investigated with the goals of: 1) reporting the level of noise contamination in each NinaPro sub-dataset, 2) proposing signal quality criteria for assessing EMG datasets, 3) analyzing the effect of signal quality on classification performance, and 4) examining the quality of the data labels.**

I. INTRODUCTION

In recent years, electromyography (EMG) signal databases have been made publicly available as benchmarking tools for the myoelectric control research community [1]–[5]. Surface EMG signals are an important source of control for several applications, including powered prostheses and human computer interaction. Myoelectric control systems exploit rich neural information contained in the EMG signals to recognize a user's movement intent and actuate a corresponding motion. The availability of benchmark datasets enables easier validation of published results, more robust comparisons between different methods, facilitates entry to the field by new researchers, and encourages contributions from other researchers in creating even larger public EMG benchmark databases. These advantages, however, can only be achieved if the quality of these benchmark databases is high. The experimental conditions must be robust, and the resulting EMG signals must contain sufficient uncontaminated neural information to produce meaningful classification results. Signal quality is particularly important for extracting meaningful gesture-related signal characteristics as the presence of noise may distort the results of feature extraction, affecting the classification results, generalizability, and even study results.

*This work was supported by MITACS, Canada

Jason Chang, Angkoon Phinyomark and Erik Scheme are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada jason.chang@unb.ca, aphinyom@unb.ca, escheme@unb.ca

Unfortunately, some amount of noise contamination is often unavoidable due to the nature of non-invasive signal acquisition from the surface of the skin. Depending on the EMG equipment and protocol, different types of noise can be found within surface EMG signals, such as motion artifact and power line interference. Moreover, the level of contamination of EMG signals in real-world environments is dynamic, which introduces confounding variability to the signals that may be unrelated to the motions of interest. Noise reduction techniques such as filtering are widely applied to suppress common noise contaminants but do not always eliminate noise completely [6]. Therefore, high quality signals are especially important in benchmarks datasets so as to obtain proper evaluations of classification performance.

The accurate labelling of data is another critical component in the evaluation of EMG-based gesture recognition as the training and testing of classifiers are dependent on the ground truth provided by the labels. Unfortunately, many factors can affect the labelling of the EMG signals, including delays from subjects reacting to a prompt, incorrect gestures (or gesture corrections mid-prompt), or early release of contractions. In such cases of mislabeling, post-hoc verification of the performed gestures is difficult, making accurate labelling essential for benchmark datasets.

The Non-Invasive Adaptive Prosthetics (NinaPro) database may currently be the largest and the most widely used EMG benchmark database to date. The NinaPro database was launched in 2014, and to date, consists of eight datasets with several different modalities, including surface EMG signals from the forearm and upper arm.

Despite the benefit and adoption of these datasets, the quality of the EMG signals in the benchmark database have not yet been formally examined. Consequently, in this work, quality of the signals in the NinaPro datasets was evaluated using six different signal quality metrics. The effect of the signal quality on the classification performance was evaluated, as was the quality of the corresponding labels.

II. METHODS

A. NinaPro Database

Surface EMG data obtained from seven NinaPro datasets (DB2-DB8), comprising over 28,400 repetitions of various hand gestures, were analyzed in this study. NinaPro DB1 was not included because it provided only a root-mean-square rectified version of the EMG signal, and not the raw signal itself. In all analyses, EMG channel 9 and 10 in DB6 were ignored as no signals were recorded from those channels.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

TABLE I: A summary of the NinaPro datasets. Electrode Configuration A: 8 equally spaced electrodes around the forearm, 2 electrodes over the flexor digitorum superficialis and the extensor digitorum superficialis, and 2 electrodes over the biceps and triceps brachii. Electrode Configuration B: Two rows of 8 equally spaced electrodes around the forearm.

Dataset	Ref.	Intact-limbed Subjects	Amputee Subjects	No. of EMG Channels	Electrode Configuration	No. of Gestures	No. of Repetitions	Equipment	Sampling Frequency
DB2	[1]	40	0	12	A	49	6	Delsys Trigno	2000 Hz
DB3	[1]	0	11	12	A	49	6	Delsys Trigno	2000 Hz
DB4	[2]	10	0	12	A	52	6	Cometa Wave Plus	2000 Hz
DB5	[2]	10	0	16	B	52	6	Thalmic Labs Myo	200 Hz
DB6	[2]	10	0	16	B	7	12	Delsys Trigno	2000 Hz
DB7	[4]	20	2	12	A	40	6	Delsys Trigno	2000 Hz
DB8	[5]	10	2	16	B	9	10	Delsys Trigno	1111 Hz

TABLE II: Signal quality metrics, with values converted to decibel (dB) using $10log_{10}(value)$. P_{signal} and P_{0-20Hz} represent the total signal power and the sum of all power densities below 20 Hz, respectively. $P_{50,100...Hz}$ is the power densities at 50 Hz and its harmonic frequencies up to the half of the sampling frequency. M_n is the *n*-th order spectral moments. $P_{upper20\%}$ is the sum of the power densities in the upper 20% frequencies (e.g, 800-1000 Hz for the EMG signals obtained with a 2000 Hz sampling frequency). *MPD* represents the mean power density of 13 consecutive points shifted across all frequencies. P_{rest} is the total signal power measured during resting.

Metric	Measurement	Equation	Acceptable level
SMR	Motion artifact (0-20 Hz)	$\frac{P_{signal}}{P_{0-20Hz}}$	>12 dB [7]
SPR	Powerline interference and its harmonics (50,100,... Hz)	$\frac{P_{signal}}{P_{50,100,...z,Hz}}$	N/A
OHM	Low and high frequency noise, particularly 20-50 Hz	$\frac{(M_2/M_0)^{0.5}}{M_1/M_0}$	< 1.4 [7]
SHR	High frequency noise (upper 20% frequency)	$\frac{P_{signal}}{P_{upper20\%}}$	> 15 dB [7]
DPR	Sufficient sampling frequency and white Gaussian noise	$\frac{MPD_{highest}}{MPD_{lowest}}$	> 30 dB [7]
SNR	Baseline noise when resting	$\frac{P_{signal}}{P_{rest}}$	See Fig. 1

Two subjects (S6 and S7) from DB3 were also excluded for the same reason.

Signals provided in the NinaPro database are pre-processed data with a notch filter (50 Hz) and a band-pass filter for DB2, DB3, DB6, DB7, DB8 (20-450 Hz), and DB4 (10-1000 Hz). Table I presents a summary of the datasets, and all the NinaPro datasets are available at http://ninapro.hevs.ch.

B. Evaluation Metrics

Six signal quality metrics were computed to quantify the level of contamination by different types of noise in the EMG signals: the signal-to-motion-artifact ratio (SMR), the signal-to-power-line-interference ratio (SPR), the power spectrum deformation ratio (OHM), the signal-to-high-frequency-noise ratio (SHR), the spectrum maximum-to-minimum drop in power density (DPR), and the signal-to-noise ratio (SNR) [7], [8]. It is important to note that although noise power in the SNR generally refers to background noise, some EMG studies have used the high frequency noise (as defined in the SHR) for SNR and so one should exercise caution when comparing SNR values in the literature. Table II presents a summary of the signal quality metrics used here, along with the reported criteria for acceptable levels.

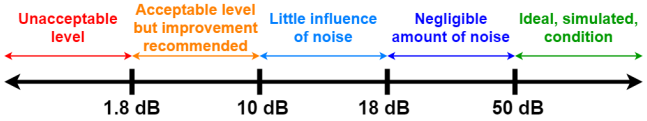

Fig. 1: Description of different SNR levels.

In this study, a taxonomy of five different SNR levels was proposed based on the literature [8]–[11] to characterize EMG signal quality, as shown in Fig 1. SNR values below 1.8 dB were considered as unacceptable (insufficient), and hence these signals should be avoided, recollected, or qualified. EMG signals with SNR values between 1.8 dB and 10 dB were deemed acceptable but improvement is recommended [9]. For SNR values between 10 dB and 18 dB, the EMG signals show little influence from noise; Phinyomark et al. [10] have shown that various feature values show less than 10% change in feature values when the SNR is above 10 dB. EMG signals over 18 dB contain negligible amount of noise, and the noise cannot be detected by either visual inspection or by classification [8], [11]. Lastly, > 50 dB SNR represent ideal conditions, typically observed only in simulated EMG signals.

To investigate the relationship between the SNR values and classification accuracy, the Pearson correlation coefficient r was computed for the six common gestures in DB2-DB7: large diameter grasp, index-finger extension grasp, medium wrap, writing tripod grasp, power sphere grasp, precision sphere grasp. DB8 was excluded in this analysis due to a lack of these common gestures.

Classification accuracies were computed using a linear support vector machine (SVM) classifier and four time domain features: mean absolute value, sign slope change, zero crossing, and waveform length [12]) with a window size of 200 ms and increments of 100 ms.

NinaPro datasets provide two sets of labels for the EMG signals: 1) a set of labels that were pre-determined from the data collection, i.e., subjects were assumed to be following the prompts on the computer screen, and 2) a set of relabelled data obtained using an offline generalized likelihood ratio method [13].

III. Results and Discussion

A comprehensive analysis of the signal quality of each NinaPro dataset was conducted. The average values for SMR,

TABLE III: Signal quality metrics obtained for each NinaPro database. The values represent average value ± standard deviation (min, median, max).

Dataset	SMR	SPR	OHM	SHR	DPR	SNR
2	21.5 ± 4.8 (9.5, 21.1, 45.2)	15.6 ± 0.7 (12.6, 15.5, 19.9)	1.3 ±m 0.1 (1.2, 1.3, 5.5)	35.3± 3.0 (23.5, 35.5, 45.3)	53.2 ± 3.1 (41.0, 53.4, 65.6)	8.6 ± 3.5 (-8.9, 8.7, 21.2)
3	18.82 ± 5.73 (1.8, 18.3, 47.9)	14.5 ± 0.6 (12.4, 14.4, 26.2)	1.3 ±m 0.6 (1.2, 1.3, 15.1)	32.3± 4.7 (22.6, 32.8, 43.3)	49.2 ± 5.1 (31.5, 49.7, 61.7)	6.4 ± 5.5 (-10.9, 5.6, 20.5)
4	21.8 ± 5.6 (11.2, 20.3, 55.8)	15.4 ± 1.4 (13.6, 15.3, 41.7)	1.4 ± 0.4 (1.2, 1.4, 18.9)	17.9 ± 3.5 (8.5, 17.8, 41.0)	34.6 ± 3.8 (23.3, 34.7, 47.7)	10.8 ± 4.0 (-2.6, 10.9, 25.1)
5	15.4 ± 2.6 (10.4, 14.9, 35.9)	20.1 ± 1.1 (16.3, 20.0, 24.2)	1.1 ± 0.01 (1.1, 1.1, 1.2)	-1.3 ± 0.4 (-2.7, -1.4, 1.0)	14.7 ± 1.7 (9.4, 14.7, 24.6)	13.1 ± 3.8 (-3.8, 13.1, 25.4)
6	17.0 ± 4.0 (3.5, 17.4, 40.6)	14.3 ± 0.7 (12.2, 14.2, 19.9)	1.4 ± 0.2 (1.2, 1.4, 3.7)	16..6 ± 0.6 (11.1, 16.5, 19.9)	39.0 ± 1.2 (32.2, 38.9, 45.4)	6.1 ± 3.2 (-9.0, 6.4, 19.8)
7	22.0 ± 4.4 (8.2, 21.3, 47.8)	14.3 ± 0.4 (11.4, 14.3, 17.0)	1.3 ± 0.1 (1.2, 1.3, 1.6)	16.7 ± 0.8 (13.3, 16.7, 21.2)	36.3 ± 1.5 (31.7, 36.1, 43.3)	9.7 ± 4.3 (-2.7, 9.5, 26.6)
8	22.9 ± 4.9 (8.6, 22.5, 45.8)	14.9 ± 0.4 (13.3, 14.8, 18.5)	1.3 ± 0.1 (1.2, 1.3, 1.5)	18.5 ± 0.8 (14.4, 18.6, 21.6)	40.0 ± 1.3 (32.9, 39.9, 47.8)	7.5 ± 4.6 (-3.2, 7.0, 24.2)

SPR, and SHR were found to be better than the acceptable levels for most of the datasets (Table III). This is mainly due to fact that filters were applied to suppress low and high frequency noise (<20 Hz and >450 Hz) as well as power line interference (50 Hz). In addition to filtering, the experimental protocol, which required subjects to perform gestures in a static limb-position while seated in a chair, likely reduced motion artifacts (<20 Hz). Although there was no effect of filtering on the frequency band of 20-50 Hz, the average OHM values were also in the acceptable range. There were trials, however, that were below the acceptable levels even after filtering, as shown by the minimum values in Table III. This indicates that the level of noise contamination in the raw EMG signals of some trials may be sufficiently high to introduce distortion of the EMG signals even after pre-processing. To maximize the quality of EMG signal, it has been recommended that "the distortion of EMG signal must be as minimal as possible with no unnecessary filtering and distortion of signal peaks and notch filters are not recommended" [14].

The average DPR values were considerably higher than the acceptable level for all datasets, except DB5 (Table III). This metric is used to ensure that the power of the EMG spectrum drops to the noise level before the upper (Nyquist) frequency limit is reached (i.e., sufficient sampling frequency). The current results are in support of a previous study [15] and suggest that using a 200 Hz sampling rate (DB5), instead of a 1000 Hz sampling rate or above (other datasets), results in a drastic reduction in discriminative information for use in EMG-based gesture recognition.

The average SNR values for DB2, DB3, DB6, DB7, and DB8 were found to be in the "acceptable but improvement recommended" range, whereas DB4 and DB5 were in the "little influence of noise" level. Lower SNR values may have been found due in part to the variability from the large number of channels and gestures. However, some negative SNR values were found (Table III), which indicate that some trials likely should be avoided due to severe noise contamination. These signals with unacceptable SNR level may be, in part, a result of incorrect labelling. It may be advisable that these signals be removed from, or qualified in, the benchmark datasets.

The common gesture sets across some of the datasets allows for the study of approximately 100 subjects, thus further investigation on SNR was performed for the six common gestures across DB2-DB7. Results showed that the SNR values for all gestures, except the writing tripod grasp gesture, was in the "little influence of noise" level (Fig. 2(a)). On the other hand, the writing tripod grasp gesture showed significantly lower SNR values ($p < 0.05$) than other gestures with a mean value of 8.3 dB. Although SNR values for all six gestures were acceptable, caution should be taken because the variability between subjects may be large due to different equipment setups (Figs. 2(b) and 2(c)).

To evaluate the effect of signal quality on classification performance, the correlation between SNRs and classification accuracies was computed using the six common gestures from 666 observations for 101 subjects. A linear positive relationship between the SNR values and classification accuracy showed a medium effect size ($r = 0.35, p < 0.05$). It should be noted that the SNR and classification performance may not have a linear relationship for the entire SNR range, especially at very low and high SNRs [8], [11]. These findings, however, support that higher SNRs facilitate more reliable discrimination of EMG signals.

The quality of gesture labels was evaluated by visual inspection. Observations indicated that both the original labels and the re-labels exhibited some problems. Fig. 3(a) shows an example of a case where a subject missed a contraction for one of the repetitions. Clearly, the original labels that were assigned by the computer prompts (assuming that subjects are closely following) did not account for the absence of the contraction. In addition, delays between the prompt and contraction onset (due to subject reaction time) can be seen in the first and third repetition. Although the delays were successfully corrected by the re-labelling method, the mislabelling for the missed repetition was not fixed.

Fig 3(b) shows what is likely to be inaccurate labelling of the no motion class. The EMG signals showed a continuation of contraction between the first two repetitions but the re-labelling method falsely detected a decay and therefor incorrectly labelled active signals as the no motion class. Features extracted from these mislabelled portions of the

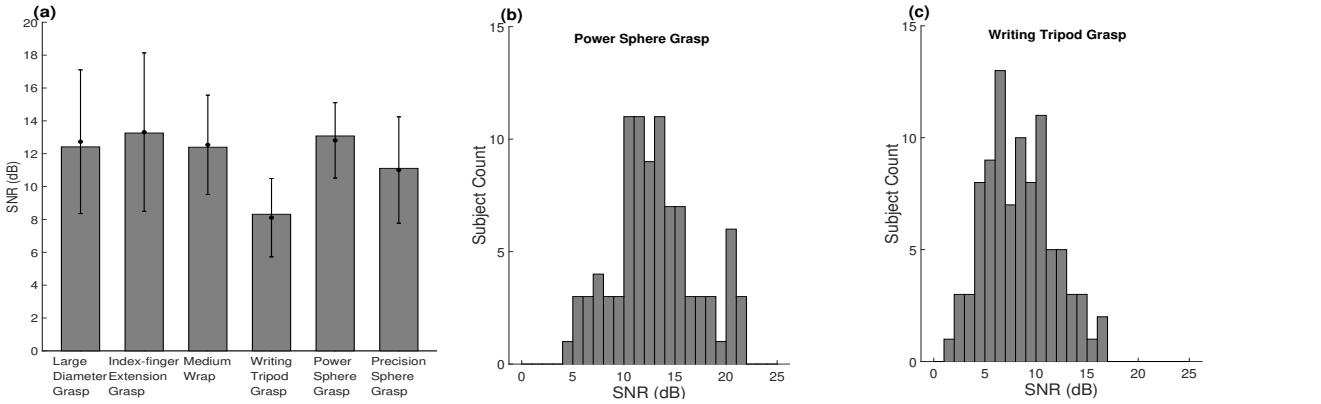

Fig. 2: SNR analysis of the gestures. SNR values represent the averaged SNR across the 5 most active channels (chosen for each dataset) and across all datasets. (a) SNR values for the six common gestures across DB2-DB7. Bars represent mean values, and dot markers represent median values with 25th and 75th percentile. (b) Histogram for the power sphere grasp gesture. (c) Histogram for the writing tripod grasp gesture.

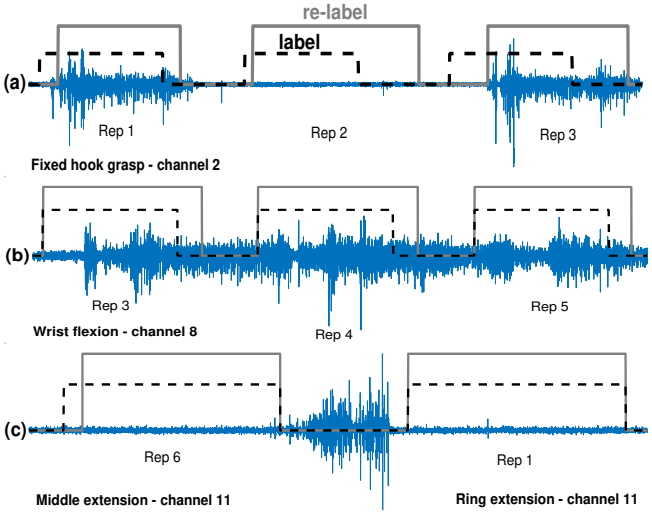

Fig. 3: Examples of incorrect labelling, (a) active class labels for a missed repetition (DB6 - subject 8, day 2, trial 1), (b) active signals labelled as no motion class (DB2 - subject 26, exercise 1), (c) failure to detect the delayed onset of the signal (DB4 - subject 2, day 2, exercise 1).

EMG signals could lead to increased feature variability and confusion in the no motion class. In Fig 3(c), the classes were completely mislabelled, and the re-labelling process did not show any improvement. It is important to note that the active signal should not be considered as a delayed version of the intended gesture. The source of the active signal is unknown, and thus it could be from a wrong motion, noise, or some other external disturbance. Such cases should therefore be discarded from further analysis.

In conclusion, a comprehensive analysis of the signal quality and labelling of NinaPro datasets was conducted. Results indicate that the signal quality of the datasets was acceptable in most cases, although some signals suffer from low SNR. More importantly, the labelling of some EMG signals was shown to be a concern throughout all datasets.

Although the NinaPro database remains a strong (and recommended) contribution to the research community, researchers should use caution when reporting results without first understanding the experimental conditions and labels.

REFERENCES

[1] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. Mittaz Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography Data for Non-Invasive Naturally-Controlled Robotic Hand Prostheses," *Scientific Data*, vol. 1, Dec 2014.

[2] S. Pizzolato, L. Tagliapietra, M. Cognolato, M. Reggiani, H. Müller, and M. Atzori, "Comparison of Six Electromyography Acquisition Setups on Hand Movement Classification Tasks," *PLOS ONE*, vol. 12, pp. 1–17, Oct 2017.

[3] F. Palermo, M. Cognolato, A. Gijsberts, H. Müller, B. Caputo, and M. Atzori, "Repeatability of Grasp Recognition for Robotic Hand Prosthesis Control Based on sEMG Data," in *International Conference on Rehabilitation Robotics*, pp. 1154–1159, July 2017.

[4] A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Improved Prosthetic Hand Control with Concurrent Use of Myoelectric and Inertial Measurements," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, July 2017.

[5] A. Krasoulis, S. Vijayakumar, and K. Nazarpour, "Effect of User Practice on Prosthetic Finger Control with an Intuitive Myoelectric Decoder," *Frontiers in Neuroscience*, vol. 13, p. 891, 2019.

[6] C. J. D. Luca, L. D. Gilmore, M. Kuznetsov, and S. H. Roy, "Filtering the Surface EMG Signal: Movement Artifact and Baseline Noise contamination," *Journal of Biomechanics*, vol. 43, no. 8, pp. 1573–1579, 2010.

[7] C. Sinderby, L. Lindstrom, and A. E. Grassino, "Automatic Assessment of Electromyogram Quality," *Journal of Applied Physiology*, vol. 79, no. 5, pp. 1803–1815, 1995.

[8] G. D. Fraser, A. D. C. Chan, J. R. Green, and D. T. MacIsaac, "Automated Biosignal Quality Analysis for Electromyography Using a One-Class Support Vector Machine," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, pp. 2919–2930, Dec 2014.

[9] Delsys, "Delsys EMGWorks." [Online] `http://www.delsys.com/emgworks/`, signal quality monitor.

[10] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "EMG Feature Extraction for Tolerance of White Gaussian Noise," in *International Workshop and Symposium Science Technology*, Dec 2008.

[11] P. McCool, G. D. Fraser, A. D. C. Chan, L. Petropoulakis, and J. J. Soraghan, "Identification of Contaminant Type in Surface Electromyography (EMG) Signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, pp. 774–783, July 2014.

[12] K. Englehart and B. Hudgins, "A Robust, Real-Time Control Scheme for Multifunction Myoelectric Control," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 848–854, July 2003.

[13] G. Staude and W. Wolf, "Objective Motor Response Onset Detection in Surface Myoelectric Signals," *Medical Engineering Physics*, vol. 21, no. 6, pp. 449–467, 1999.

[14] M.B.I. Reaz and M.S. Hussain and F. Mohd-Yasin, "Techniques of EMG signal Analysis: Detection, Processing, Classification and Applications," *Biological Procedures Online*, vol. 8, pp. 11–35, 2006.

[15] A. Phinyomark, R. N. Khushaba, and E. Scheme, "Feature Extraction and Selection for Myoelectric Control Based on Wearable EMG Sensors," *Sensors*, vol. 18, no. 5, 2018.