Optimized Algorithm for Muscular Diseases Recognition Based On Temporal Parameters Analysis and Correlation Coefficients

E.A. Mejia-Gonzalez, J.A. Lopez-Leyva, J. Estrada-Lechuga and M.A. Ponce-Camacho
CETYS University, Centre of Innovation and design Ensenada, Baja California, Mexico
josue.lopez@cetys.mx

Abstract. In this paper, an optimized algorithm based on temporal parameters analysis and correlation coefficients is presented in order to perform muscular diseases recognition. Statistical information was measured of three classes signals (Healthy, Myopathy and Neuropathy conditions). The temporal parameters that were initially proposed (14) were optimized based on the correlation coefficients. Thus, only 9 parameters were selected for optimized the algorithm, and the time required for training and recognition is \( \approx 0.2s \) and \( \approx 4ms \), respectively.

1. Introduction
Nowadays, the bio-mechatronic and bio-informatics fields have had a huge technological development with important performance improvements. In particular, the gesture and medical conditions recognition for many parts of the body (e.g. shoulder, knee, face, arm, among others) using high-end computational resources permit to propose a wide variety of applications for the industrial and health sector [1, 2]. All of these required a database (sometimes called feature vector) of important features to analyse different patterns and determine the maximum likelihood related to the data previously recorded. The feature vector collective value of parameters (which are in the temporal or/and frequency domain) that represent a particular signal in order to determine a particular disease [3]. However, sometimes a large number of parameters and techniques are used in order to increase the efficiency of an algorithm, which commonly means increase the computation capabilities [4]. In particular, there exist many technical proposals for monitor and analysis electromyogram (EMG) signals. In fact, many of them are based on the superficial technique (using electrodes) called Surface EMG (sEMG), while others based on invasive action or muscle biopsy (using a needle). These proposals use the Bayesian method and Markov models in order to determine the gestures. In some cases, iterative optimization is performed [5, 6]. Also, because a particular gesture or condition usually is related to multiple muscles, a multi-category classification problem is presented for complex activities or movements [7]. In addition, the cost of high-end technology with potential applications in the design of prostheses makes it difficult to use widely by society. In this way, the design of software and hardware considering the minimum requirements with optimized performance and high processing speed is necessary. Finally, based on the aforementioned, a high-speed algorithm optimized for biomedical condition recognition is proposed. This paper is organized as follow. Section 2 shows the theoretical background regarding the statistical analysis of biosignals and optimization process of the algorithm proposed. Next, Section 3 presents the numerical
results respect the classification, statistical analysis, and optimization. Finally, Section 4 shows the conclusion.

2. Theoretical Background

The raw \((R_{i,j}(t))\) data for three different classes \((i)\) is used, where \(j\) is a particular signal (up to 20 training signals). Next, for each \(R_{i,j}(t)\) performing a feature engineering in order to calculate 14 (parameters \(b\)) features such as: median, RMS value, variance, kurtosis, skewness, interquartile range (IQR), quartile 1 (Q1), quartile 2 (Q2), quartile 3 (Q3), 10th percentile, 20th percentile, 50th percentile, 70th percentile, and 90th percentile. These parameters permit to training the algorithm to recognized a particular class based on a specific feature vector \((f_i)\) (see (1)). In particular, each feature is possible defined as a random variable.

\[
f_i = [m_i, RMS_i, Q1_i, Q2_i, Q3_i, p10_i, p20_i, p50_i, p70_i, p90_i]
\] (1)

Once the feature vector for each class is calculated based on the use of the training signal, the algorithm uses a random testing signal \((T_{i,j}(t))\) and dynamic thresholds to determine the probability of belonging to each test signal to a particular class, i.e. estimation via maximum likelihood (± 20 %, ±10 % and, ±5 %) is performed. In particular, the algorithm makes a binary decision related to the threshold established; if the feature is within the threshold range, a true decision (high) is stored.

\[
f_{i,a} = \begin{cases} 1 & f_{i,a} = f_{r,a} \\ 0 & f_{i,a} \neq f_{r,a} \end{cases}
\] (2)

So, (3) describes the aforementioned.

\[
p_i = \left( \sum_{a=1}^{b} f_{i,a} \right) / b
\] (3)

It is important to clarify that, at the begging, the algorithm does not determine o recognize each disease for a particular class, rather, it provides statistical information to make a smart decision about that. At this moment, statistical information for each testing signal that represent a disease of a particular class is obtained based on the feature vector \((f_i)\). However, an optimized feature vector \((f_{i-opt})\) is proposed to improve the performance of the algorithm related to the statistical information. To do that, the correlation coefficient matrix \((\rho_{XY})\) is calculated between all the features for each class to determine the dependence based on statistical information. In particular, the correlation coefficient matrix of the Class 1, 2 and 3 are \(\rho_{XY}(H)\), \(\rho_{XY}(M)\) and \(\rho_{XY}(N)\), respectively. Next, an overall coefficient matrix \((\text{see (4))}\) is calculated considering the statistical independence between classes.

\[
\rho_{XY}(H, M, N) = \prod_{H,M,N}(\rho_{XY})
\] (4)

\[
|\rho_{XY}(H, M, N)|_{\text{max}} \rightarrow f_{i-opt}
\] (5)

In fact, (5) has a relation to the maximum values for each \(\rho_{XY}\), i.e., \(|\rho_{XY}(H)|_{\text{max}}, |\rho_{XY}(M)|_{\text{max}}\) and \(|\rho_{XY}(N)|_{\text{max}}\). In this manner, the probability of a correct classification is maximized \((p_i)_{\text{max}}\) using \(|\rho_{XY}(H, M, N)|_{\text{max}}\). Due to the optimization process, the parameter \(b\) will be decreased, because some particular features were deleted based on the minimum correlation values.

\[
\rho_{XY}(H, M, N) = \prod_{H,M,N}(\rho_{XY})
\] (6)

\[
|\rho_{XY}(H, M, N)|_{\text{max}} \rightarrow f_{i-opt}
\] (7)

3. Measurements and Results

This section shows the classification of raw signals that represent different biomedical conditions. Next, a feature vector is generated to represents the principal characteristics of each biomedical condition. All the analysis was performed using an Intel® Celeron® processor at 2.16 GHz. Figure 1, 2 and 3, show the biosignals for three different conditions: Healthy (Class 1), Myopathy (Class 2) and Neuropathy (Class 3), respectively. In particular, 20 signal was measured with 2000 samples per second for each
signal. These signals were obtained from the public database that archive well-characterized digital recordings of physiologic signals and related data for use by the biomedical research community. (https://physionet.org/physiobank/database/emgdb/). In fact, some different features between the classes are clearly observed. In particular, the maximum and minimum amplitude values are an important feature for Class 3, while the speed of the biosignals is increased for Class 2.

![Figure 1. EMG signals for a healthy person](image1)

![Figure 2. EMG signals for Myopathy.](image2)

![Figure 3. EMG signals for Neuropathy](image3)

In order to clarify, Figure 4 shows the best-representing signals each biomedical condition. Making a comparison between the signal presented in Figures 1, 2, 3 and 4, it is possible to observe particular characteristics.

![Figure 4. Best signals that represent EMG conditions (healthy, myopathy, and neuropathy).](image4)

Thus, 14 parameters were chosen in order to define the characteristics (i.e. feature vector) of each condition (classes) analysed. Figures 5, 6 and 7 show the particular values for parameters considering
each condition and the margin of error allowed (the thresholds are only representative, the magnitude showed does not is consider for the algorithm). These value of parameters were calculated considering the entire collected signals shown in Figure 1, 2 and 3 for each biomedical condition. Thus, some values of the parameters of different classes are similar, but others are highly different. For example, the obliquity of Class 1 is $\approx 7.7$, while for class 2 and 3 are $\approx 9.5$ and $\approx 60.3$, respectively. It is important to clarify that, at this moment, all the temporal parameters were considered for the algorithm for performing the classification and biomedical recognition. In fact, these feature vectors are recorded by the database and they are the core information in the training stage of the algorithm based on the raw signals. In addition, some features are highly similar between classes, so, a low performance based in statistical analysis will be expected. This is not a mistake, in fact, it is a technical argument that will be used for the optimization.

![Figure 5](image5.png) **Figure 5.** Characteristics of the overall EMG signal for a health condition.

![Figure 6](image6.png) **Figure 6.** Characteristics of the overall EMG signal for a Myopathy condition

![Figure 7](image7.png) **Figure 7.** Characteristics of the overall EMG signal for a Neuropathy condition

After that the characteristics vector is defined for each class, the algorithm proposed is tested in order to determine the performance. Thus, a random selection of class and biosignal is used as input to the algorithm for different error threshold, $\pm 20\%$, $\pm 10\%$, and $\pm 5\%$. These thresholds were established only for technical objectives, however, the threshold range has to be reduced for real biomedical applications. Next, an analysis of the performance for each class is performed based on a statistical population using a box plot tool. Figure 8 shows the performance of the algorithm considering only a random input signal of Class 1 and the error threshold of $\pm 20\%$. Thus, for the Class 1 performance, the median is 50\%, the first quartile is $\approx 21.4\%$, the third quartile is $\approx 64.2\%$, the maximum value is $\approx 78.8$, minimum value and,
a wide interquartile range (IQR) ≈ 42.8%. Moreover, for Class 2 and 3 performances, the median (17.8% and 7.1%, respectively) and IQR values decreased. Next, Figure 9 shows the performance considering a random input signal of Class 2. It is possible to see that the maximum percentage of recognizing signal (≈85.7%) is in Class 2. Also, Figure 10 shows the performance of the algorithm for Class 3, ≈85.7%.

Next, the error threshold is reduced to ±10% and, the performance of the algorithm was measured. Figure 11, 12 and 13 show the statistical performance for the random input signals of classes 1, 2 and 3, respectively. Here, the principal variation is related to the median reduction in comparison to the results considering an error of ±20%. However, the classification and recognition processes are still adequate. The principal argument of the mentioned is that the statistical information for each class is congruent with the input signal used for testing. For example, when the input signal of Class 1 was used, the maximum probability corresponds to Class 1, while the other classes have less occurrence probability. This statistical scenario is the same for the input signal of Class 2 and 3. Next, the performance of the algorithm was measured with a lower error threshold, ±5% with the conditions aforementioned. Figure 14, 15 and, 16 show the statistical information related to each particular class. Finally, considering the statistical performance for all the classes and thresholds presented, it is possible to observe that the classification and recognition process is correct under certain statistical considerations, i.e. the maximum probability always was the correct regarding a certain class, although the value of the
probability is not maximized. These means that the median value of the probability for all the classes was reduced according to the decreasing of threshold (i.e. from ±20% up to ±5%).

Figure 11. Boxplot of the statistical performance of the algorithm for Class 1 considering a tolerance of ±10%.

Figure 12. Boxplot of the statistical performance of the algorithm for Class 2 considering a tolerance of ±10%.

Figure 13. Boxplot of the statistical performance of the algorithm for Class 3 considering a tolerance of ±10%.

Figure 14. Boxplot of the statistical performance of the algorithm for Class 1 considering a tolerance of ±5%.

Figure 15. Boxplot of the statistical performance of the algorithm for Class 2 considering a tolerance of ±5%.

Figure 16. Boxplot of the statistical performance of the algorithm for Class 3 considering a tolerance of ±5%.

As was shown, the statistical information related to the classification and recognition of the biomedical condition is adequate based on the results. Now, an optimized feature vector is calculated based on the overall correlation coefficients between all classes to use it by the algorithm. In order to clarify, the results showed were obtained based on the 14 features mentioned, i.e. 14 particular values related to
each feature for each biomedical condition. Firstly, the correlation coefficients for each class are calculated. Figure 17, 18 and 19 show the correlation coefficients values of each class (i.e. Healthy, Myopathy, and Neuropathy, respectively). Because the algorithm proposed have to recognize a single input signal related to each class, the particular correlation values of each class have to be processed in order to determine the principal temporal parameters with highest correlation values. Figure 20 shows the final coefficients values of correlation between all the classes analysed. The coefficients in the diagonal are not useful, due to they represent the autocorrelation.

**Figure 17.** Correlation coefficients for Class 1.

**Figure 18.** Correlation coefficients for Class 2.

**Figure 19.** Correlation coefficients for Class 3.

**Figure 20.** Correlation coefficients of all classes.

**Figure 21.** Statistical results based on the optimized algorithm
Hence, based on the results shown in Figure 20, 9 features were selected to permit the optimization of the overall feature vector for detecting all the classes in an improved way. Using this optimized feature vector, the performance of the algorithm was measured again. Figure 21 shows the statistical information using a testing signal of Class 1. As can be seen, all the statistical parameters were improved; the median of Class 1 is to ≈75% (while it was ≈50% using a non-optimized feature vector, see Figure 8), and the median of the probability of the other classes decreased. Finally, Figure 22 shows the time required by the algorithm for training and recognition stages, 200 and 4 milliseconds average, respectively, which means an important experimental result for future applications.

4. Conclusion
The performance of an optimized algorithm for classifying biomedical conditions based on statistical analysis of EMG signals is presented. The optimization was performed based on the correlation coefficients between all the biomedical conditions in order to reduce the temporal parameters analyzed and considered in the initial feature vector. In this first approach, the classification performance is adequate, i.e. for all the conditions and thresholds used, the class recognized was successful, which means that the correct class presented the maximum probability. However, it is important to maximize some statistical parameters to improve the overall performance considering the real biomedical applications. In fact, as future work, frequency parameters will be added to the feature vector and a neural network will be used in order to improve the performance without increasing the computational requirements and the response time.

References
[1] He J, Sheng X, Zhu X, Jiang C and Jiang N 2019 IEEE Trans. on Ind. Inform., 15, 1226
[2] Kosmidou V E, Petrantoniakis P C and Hadjileontiadis L J 2011 IEEE Trans. on Sys., Man, and Cyber., Part B, 41, 1531-43
[3] Phinyomark A, Phukpattaranont P and Limsakul C 2012 Exp. Sys. with Appl. 39, 7420-31
[4] Wu D, Pigou L, Kindermans P-J, Le N, Do-H, Shao L, Dambre J and Odobez J-M 2016 IEEE Trans. on Patter. Analyst. and Mach. Intell., 38, 1583-97
[5] Wheeler K R, Chang M H and Knuth K H 2006 IEEE Trans. on Sys., Man, and Cyber., Part C, 36, 503-14
[6] Kosmidou V E and Hadjileontiadis L J 2009 IEEE Trans. on Biomed. Eng., 56, 2879-90
[7] Naik G R, Kumar D K and Jayadeva IEEE Trans. on Inform. Tech. in Biomed. 14, 301-08

Acknowledgments
We thank to CETYS University, Centre of Innovation and Design and National Council of Science and Technology of Mexico for the administrative and technical support.