IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication november 12, 2025, date of current version february 11, 2026.

Digital Object Identifier 10.1109/ACCESS.2025.50829

Grouping sEMG signals from individuals with neurological injuries using clustering methods

JOÃO PEDRO M. LOURENÇÃO[1], Heitor S. Lopes[1], María Verónica G. Méndez[1], Cristian Veggian Matias [1], Daniel P. Campos[2], and José Jair A. Mendes Júnior[1]

[1]Postgraduate Program in Electrical Engineering and Industrial Informatics, UTFPR, Curitiba, PR 80230-901 Brazil
[2]Postgraduate Program in Biomedical Engineering, Curitiba, PR 80230-901 Brazil

Corresponding author: João Pedro M. Lourenção (e-mail: jlourencao@alunos.utfpr.edu.br).

This work was supported by the National Council for Scientific and Technological Development (CNPq - Brazil).

ABSTRACT The contamination of surface electromyography (sEMG) data by artifacts is a critical issue, particularly when acquired from individuals with neurological injuries. This necessitates subjective and time-consuming visual inspection to verify signal utility. To address this, this work proposes an automated quality assessment method based on unsupervised learning to distinguish usable signals from those compromised by artifacts, such as electrode detachment or saturation. Features were extracted using the Time Series Feature Extraction Library (TSFEL), and the top 80 discriminative features across temporal, statistical, and spectral domains were selected via ANOVA. K-means and Agglomerative Clustering were employed to group the data, followed by an Isolation Forest stage to refine the selection of usable signals. Validation against an objective ground truth revealed the complexity of the task (ARI ≈ 0.20), yet the pipeline effectively retained 52% of the dataset as usable signals. Furthermore, a comparison between the retained injured signals and healthy controls demonstrated statistically significant differences ($p < 0.001$) with a medium-to-large effect size (Cohen's $d \approx 0.72$), confirming that the automated approach preserves signals with distinct and biologically relevant spectral characteristics.

INDEX TERMS Clustering methods, Feature ablation, Neurological Injuries, Surface Electromyography

I. INTRODUCTION

A Global Burden of Disease study in conjunction with the World Health Organization has shown that nervous system disorders are the leading cause of disability-adjusted life years (DALYs) worldwide, affecting 3.4 billion people and accounting for 443 million DALYs in 2021. Age-standardized mortality rates and DALY rates decreased by 33.6% and 27%, respectively, between 1990 and 2021. The main conditions contributing to this burden include stroke, neonatal encephalopathy, and Alzheimer's disease [1]. In Brazil, where this study was conducted in 2023, 8.9% of the population had some disability, with 1.4% facing difficulties in performing manual tasks, such as handling items or operating containers [2].

Myoelectric signals (electromyography, EMG), more specifically, surface electromyography (sEMG), have been used in several areas, including rehabilitation and physiotherapy [3]–[5], control of prosthetic devices [6], [7], assessment of muscle fatigue [8], [9], and motion recognition [10], [11]. The sEMG technique enables the non-invasive acquisition of muscle signals, ensuring that patients experience no pain or discomfort. Combined with sEMG, functional electrical stimulation (FES) acts as a feedback mechanism. It is becoming increasingly common to apply electrical stimulation in stroke cases as a temporary aid in the rehabilitation process, to assist motor learning and functional recovery [12], [13]. FES allows people with neurological injuries to move underutilized limbs, preventing muscle atrophy and providing more fluid and natural movements [14].

However, sEMG is significantly impacted by several types of noise, including the baseline noise, which is generally modeled as Gaussian white noise; electrochemical noise, caused by the impedance of the tissue in contact with the electrode; baseline wander, caused by body movements; interference, from several origins, such as the electrocardiogram; crosstalk from other muscles and electromagnetic (system power supply). In addition, movement artifacts can be present, since they are disturbances caused by body or cable movements, and can generate peaks or fluctuations in the signal [15].

Acquiring sEMG from people with neurological conditions poses additional challenges. These individuals often have

increased difficulty performing movements, which increases the system's vulnerability to noise and can lead to data loss [16]. Data loss detection occurs after a delay, identified by experts or post-processing methods [15], since data acquisition equipment usually lacks pre-processing and quality checking. Identifying the nature of the signal, usable or unusable, at the acquisition time becomes important to warn who is in charge of acquiring the signals if there is a need to re-collect data, therefore avoiding the need for extra days of acquisition and loss of information.

Figure 1 illustrates the problem, where a health agent in a clinical environment collects sEMG signals and needs to assess their usability. This is difficult without advanced post-processing, mainly because current acquisition systems often lack a visual display. Therefore, a key innovation would be the ability to immediately check the signal's quality just after collection and, without performing a full classification, differentiate between signals from individuals with muscle injuries and those who are healthy. This would lead to a faster and more intuitive evaluation of both signal quality and clinical relevance.

Aiming to address these issues, this study proposes a quality assessment method for sEMG signals acquired from individuals with neurological conditions in a hospital environment. The primary objective is to identify usable signals for data processing through clustering algorithms, eliminating those affected by noise and information loss. As a secondary analysis, the cleaned dataset is compared with signals from healthy individuals to demonstrate physiological differences in frequency and amplitude domains. All processed data will be made publicly available via an online repository. The contributions of this work are: (i) the development of a hybrid unsupervised pipeline that validates K-Means against Agglomerative Clustering and employs Isolation Forest for refinement to recover physiologically valid signals often discarded by rigid thresholding; (ii) a systematic feature selection analysis identifying the top 80 temporal, statistical, and spectral features most discriminative for sEMG quality; and (iii) the publication of a challenging, real-world sEMG dataset collected from individuals with neurological injuries using low-cost hardware.

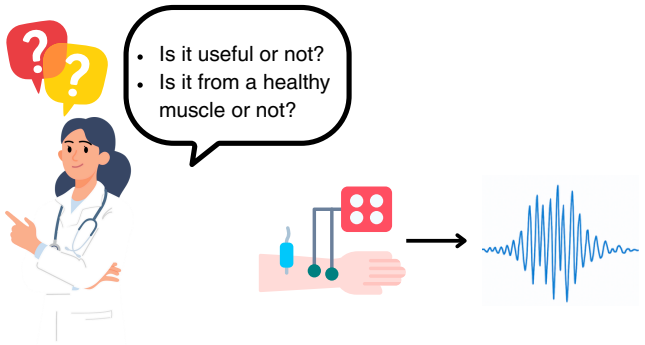


FIGURE 1: Problem illustration - A healthcare professional wonders whether the signals just collected are useful or not, and whether the patient is healthy or not.

II. MATERIALS AND METHODS

In this Section, we present the methodology used for collecting and processing signals from volunteers, as shown in Figure 2.

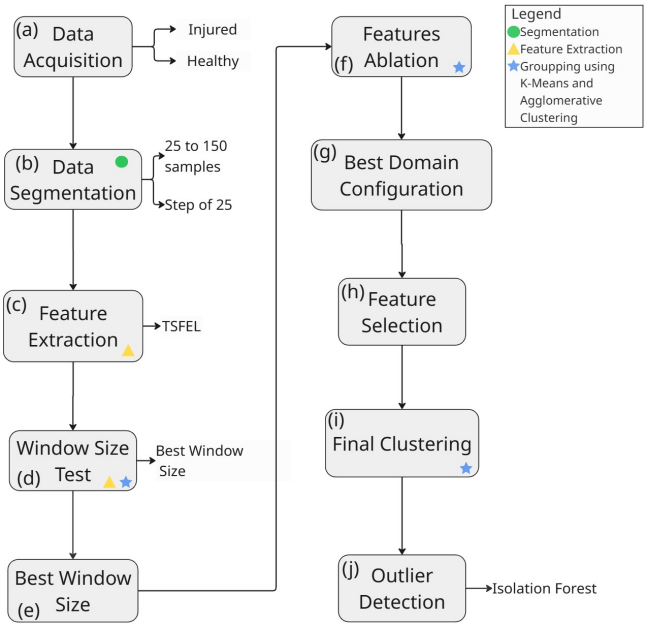


FIGURE 2: Methodology flowchart applied in this work.

A. DATA ACQUISITION

Data was acquired from two groups of volunteers: healthy and injured.

- **Healthy volunteers**: were individuals without myopathy or neurological diseases who could perform manual tasks. Their data was collected at the Federal University of Technology – Paraná, with approval from the Research Ethics Committee (process nº 7.218.395).
- **Injured volunteers**: were subjects with neurological diseases, such as post-stroke and cerebral palsy. Their data was collected at the Pequeno Cotolengo Health Complex, with approval from the Research Ethics Committee of Londrina State University (process nº 7.047.015).

The acquisition hardware consists of an ESP32 microcontroller that receives and stores data on an SD card, and a biopotential acquisition circuit based on the AD8232. The circuit for acquiring the sEMG signals, presented in Figure 3, was developed based on the works of [17] and [18]. It receives data from electrodes and sends it to the microcontroller. Each signal was collected for approximately 45 s, at a sampling rate of 1 kHz.

The sEMG signals were acquired from the *triceps brachii* and *extensor carpi radialis*, with the electrodes placed as shown in Figure 4. The reference electrode was placed at

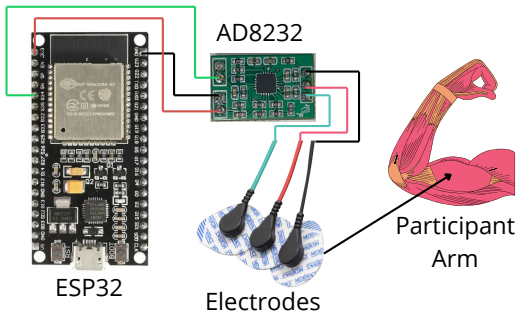


FIGURE 3: Hardware used for data acquisition.

the bony protrusion of the elbow. For *triceps brachii*, the electrodes were placed around two to three fingers above the reference, as shown in Figure 4 (a). The electrode on the *extensor carpi radialis* was placed on the forearm, three fingers below the reference, as shown on Figure 4 (b).

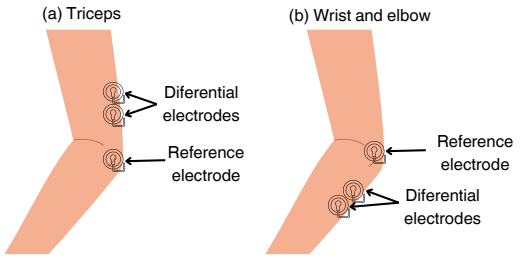


FIGURE 4: Electrodes positioning for collecting data: (a) on the triceps, (b) on the wrist and elbow.

The acquisition protocol was based on three approaches. The first is the wrist evaluation (A), which consists of opening the hand for five seconds, resting for three seconds, and another five-second extension. The second approach is a reach evaluation (B), consisting of an arm extension as far as possible for five seconds, rest (three seconds), and another five-second extension. Finally, the last approach is the elbow evaluation (C), which is an extension (in the most challenging position) for five seconds, rest (three seconds), and another five-second extension.

The data from the injured people was acquired from five volunteers, as shown in Table 1. The dataset used in this work is related to the development of [19]. This procedure involved the use of electrical stimulation in injured volunteers. This dataset was acquired using the same circuit previously described. The dataset [1] is organized into two labels, "pre" and "pos". The first refers to the movements done before applying the FES, and the second refers to the movements after the FES application, as shown in Table 1. The number beside the letter "P" refers to the patient's number. A total of 158 files were acquired from five injured volunteers.

The data acquisition from the healthy individuals was performed using the same protocol as used for the injured volunteers. For each healthy subject, two signals were acquired for each movement, resulting in 24 healthy signals. These signals were used to balance the database.

TABLE 1: Data for each injured patient

Volunteer	PreA	PreB	PreC	PostA	PostB	PostC	Total
P1	9	10	10	10	9	10	58
P2	0	1	1	1	1	1	5
P3	7	4	7	10	9	9	46
P4	9	-	6	9	-	6	30
P5	1	1	1	1	1	1	6
Total							158

Figure 5 shows examples of what are considered unusable signals. These signals do not contain information for later analysis and may be composed entirely of noise or saturation. Figure 6 shows examples of signals considered usable for later analysis. We can see that even if they have some data loss, it is not enough for the signal to be discarded, as it does not make it completely unusable.

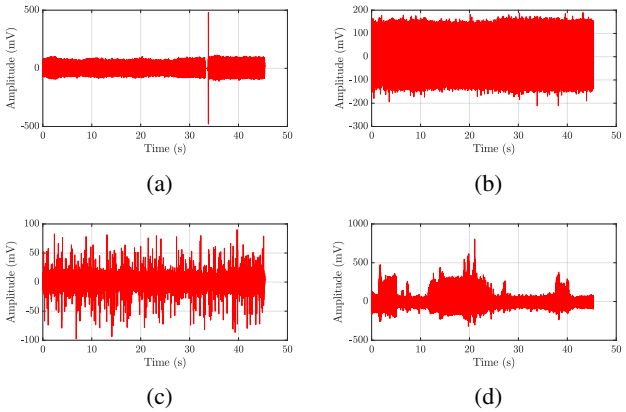


FIGURE 5: Representative examples of signals that can be considered unusable. (a) Signal contaminated by power line interference (60Hz) with transient spikes. (b) Signal dominated by pure power line interference (60Hz). (c) Signal exhibiting saturation. (d) Signal corrupted by motion artifacts.

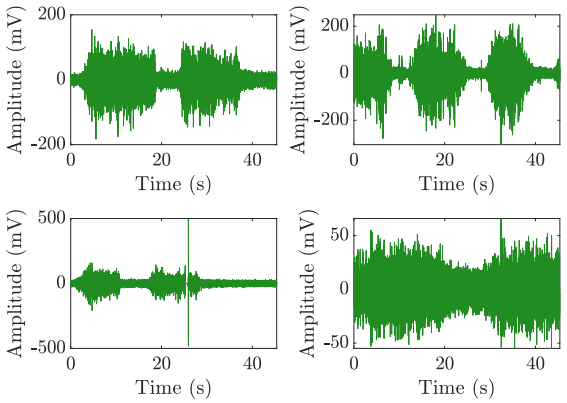


FIGURE 6: sEMG signal examples that can be considered usable.

[1] https://doi.org/10.5281/zenodo.15850286

To establish a rigorous ground truth for validating the unsupervised methods, we defined quantitative criteria to categorize signals as ''Unusable''. As standard consensus values for artifact rejection are currently established as a persistent challenge in the field [20], the specific thresholds below were empirically determined based on the dynamic range and noise floor of our acquisition hardware:

- **Saturation:** Signals where the amplitude reached the hardware limits (0 or 4095) for more than 1.0% of the recording duration.
- **Signal-to-Noise Ratio (SNR):** Signals with an estimated SNR below 6.0 dB are considered critical for distinguishing low-level muscle activity from background noise in clinical environments.
- **Electrode Detachment:** Periods where the rolling standard deviation dropped below 5 μV or more than 500 ms, indicating loss of skin contact.

Based on these objective criteria, 39.9% of the dataset (63 out of 158 signals) was classified as unusable. This Ground Truth vector was subsequently used solely for external validation of the unsupervised results.

B. SIGNAL PROCESSING AND FEATURE EXTRACTION

The first step of signal processing is filtering the signals with a 60 Hz notch filter and a 4th-order Butterworth bandpass filter, which had a low cut frequency of 20 Hz and a high cut frequency of 450 Hz [21]–[24]. This was done to remove electromagnetic interference and limit the frequency range of the sEMG signals.

For the initial test, we varied the signal window length from 25 to 150 samples, in steps of 25, to define the signal range for feature extraction (see Figure 2 (b)). We also analyzed segmentation with and without overlap, with the overlap defined as 50%. This approach allows for analysis using a minimal number of samples while achieving results similar to those with a higher number of samples. Additionally, it enables the identification of local patterns, noise reduction, greater precision and control, and saves time and resources when analyzing large datasets [25], [26].

The raw sEMG signals were processed to extract relevant features for quality classification. Feature extraction was performed using the Time Series Feature Extraction (TSFEL) Library [27] (Figure 2 (c)). Initially, a comprehensive set of features from Temporal, Statistical, and Spectral domains was extracted to capture the diverse nature of signal artifacts.

To objectively determine the optimal number of attributes for quality control, we performed a systematic feature selection analysis. First, features were ranked based on their discriminative power using the Analysis of Variance (ANOVA) F-value method. Subsequently, we iteratively evaluated the anomaly detection performance (measured by the Adjusted Rand Index against the ground truth) for feature subsets varying from k=5 to k=150 features. This analysis revealed that the performance peaked at a subset of 80 features. This optimized subset predominantly included temporal and statistical metrics (e.g., Absolute Energy, Histogram Entropy, Interquartile Range), which effectively capture the high-amplitude artifacts (saturation) and low-variability periods (sensor detachment) that characterize unusable data.

After determining the minimum number of samples and the best feature subset, the clustering was performed to group the signals using two methods: k-means and Agglomerative Clustering, varying the number of clusters from 2 to 7 (Figure 2 (h)). By using the silhouette score, the optimal number of clusters was found.

The t-SNE method [28] was used to visualize the signals grouped in a hyperplane projected into a two-dimensional space. Then, the results of the clustering methods were analyzed and compared with the Isolation Forest [29] (Figure 2 (i)), which is an unsupervised machine learning algorithm capable of identifying anomalies or outliers in the data. The contamination parameter for the Isolation Forest was set to 0.4, selected to align with the prevalence of unusable signals observed in the ground truth analysis (approx. 40%). By anchoring the contamination rate to the statistically observed noise proportion, we avoid arbitrary thresholding.

After discarding irrelevant data with the Isolation Forest method, the volunteers' data with and without injuries were compared to check for eventual differences between them. For instance, differences in signal amplitude and high-frequency components could be noticed when comparing these two signals. Frequency features, such as median frequency, mean frequency, and frequency ratio, were calculated to compare these signals. Healthy signals are expected to have higher values than injured signals. Lesion signals are anticipated to have smaller amplitudes and predominantly lower-frequency components [30]. This analysis would be useful for visualizing the effects of neurological injuries on muscles.

III. EXPERIMENTS AND RESULTS

In this section, we present the results, which include the window size investigation, feature extraction, and the ablation and clustering procedures. Also, a comparison between healthy and injured individuals is included.

A. WINDOW SIZE TEST

The minimum number of data points in each signal segment was determined based on the optimal clustering of the signals. This was accomplished in two ways: one method involved no data overlap, while the other allowed for overlap. For the method without overlap, we developed a function that takes a signal and a window size as inputs and returns a list of all the generated windows. Then, features were extracted for all of these windows, the average was calculated, and the result was inserted into a feature data frame, which was passed as input to the clustering methods.

Silhouette Score was calculated to compare the different window sizes. For this, a sliding window approach with a 50% overlap was employed. Subsequently, both grouping methods

were used to assess the window lengths, and the results are shown in Table 2.

TABLE 2: Silhouette scores for each window length.

	K-Means		Agglomerative Clustering	
Window Size	Without Overlapping	With Overlapping	Without Overlapping	With Overlapping
25	0.6774	0.6426	**0.9221**	0.6768
50	0.7167	0.6211	0.9220	0.7077
75	0.7196	0.5919	0.9220	0.7129
100	0.7252	0.6164	0.9220	0.7159
125	0.7144	0.5822	0.9220	0.7159
150	0.7149	0.5997	0.9220	0.7074

We observed that using overlapping windows yields nearly identical results for K-Means clustering but worse performance for Agglomerative Clustering. Based on this analysis, and after evaluating various window sizes, we selected a window size of 25 data points. This can be attributed to the inherent characteristics of each algorithm. K-Means, which minimizes intra-cluster variance by assuming spherical distributions, is robust against the temporal redundancy inherent in overlapping subsequence clustering [31]. Its performance remains stable even without complex initialization variants, making it computationally efficient for short time-series segments. In contrast, Agglomerative Clustering, as a hierarchical method that relies on local similarity, is more susceptible to this redundancy. The overlap can lead to the formation of less representative clusters, ultimately degrading their results [32].

B. PERFORMANCE VALIDATION AND GROUPING

To rigorously validate the proposed grouping method, we computed the Adjusted Rand Index (ARI) against the objective ground truth established in Section II. We compared the K-Means clustering algorithm against the Isolation Forest (anomaly detection) to determine if the grouping paradigm was the limiting factor.

The results showed comparable performance between K-Means (ARI = 0.2) and Isolation Forest (ARI = 0.24). While Isolation Forest showed a slight advantage, the similarity in results suggests that the challenge in separating usable/unusable signals in this population is not algorithm-dependent.

It is important to note that the ARI measures the agreement between the unsupervised grouping and a *rigid* set of objective criteria (e.g., SNR $>$ 6dB). In individuals with neurological injuries, paretic muscles often generate signals with naturally low amplitude and irregular firing rates [33], [34], which objective metrics may classify as ‘‘noise’’ (low SNR), whereas unsupervised methods correctly identify their spectral physiological signature. Thus, the moderate ARI values likely reflect the discrepancy between rigid engineering thresholds and the biological variability of pathological sEMG, rather than a failure of the clustering capability.

A total of 158 signals were acquired from 5 injured patients (Table 1). The clustering results, with K-Means (k=2), gave the optimal clustering (silhouette score = 0.904), with Cluster 0 (largest) retaining 136 signals, and Cluster 1 retaining 22 signals. After that, the outlier removal step, with Isolation Forest (contamination = 0.4), removed 54 signals from Cluster 0. The total final dataset becomes 82 usable signals ($\approx 62\%$ retention). Table 3 summarize the results.

TABLE 3: sEMG Quality Assessment Pipeline

Pipeline Step	**Cluster 0 (Potential Usable)**	**Cluster 1 (Unusable)**	**Total**
Total Acquired Signals	–	–	**158**
K-Means Clustering (k=2)	**136**	**22**	158
Isolation Forest Removal	-54	0	-54
Final Usable Signals	**82**	**0**	**82 (51.9%)**

Table 4 presents the confusion matrix comparing the proposed unsupervised pipeline against the objective ground truth. The system achieved an accuracy of 68.35%. Notably, a discrepancy was observed in the False Positive rate (18 signals). A post-hoc visual analysis of these specific signals revealed that the objective ground truth, based on rigid thresholds (e.g., saturation $> 2.5\%$), tends to be overly conservative. The proposed clustering pipeline successfully retained these signals, which, despite presenting marginal saturation, preserved clear motor unit action potential patterns suitable for clinical interpretation.

TABLE 4: Confusion Matrix: Automated Pipeline vs. Objective Ground Truth

Classification		**Ground Truth (Objective Criteria)**	
		Usable	**Unusable**
AutomatedPipeline	**Usable**	**64**	**18**[a]
	Unusable	**32**	**44**
Accuracy		**68.35%**	

[a] Signals with marginal artifacts recovered by the clustering method.

On the other hand, the pipeline exhibited a conservative behavior regarding False Negatives (32 signals discarded), prioritizing the purity of the dataset over quantity. In clinical applications involving pathological sEMG, where artifacts can lead to misdiagnosis, a conservative filtering approach is often preferable to ensure that subsequent analyses are based on high-confidence data.

From this, we can observe the organization cluster 0 (136 signals) in the hyperspace using t-SNE. To better visualize the signal distribution in this hyperspace, we first removed outliers using Isolation Forest. Figure 7 shows the final organization of cluster 0, with the outliers highlighted in red.

We applied the same process to agglomerative clustering. As shown in Figure 8, which illustrates the t-SNE hyperspace with outliers highlighted, it was evident that the usable signals were grouped identically by both clustering methods.

Considering the 158 signals collected, both methods classified 82 as usable. This consistent grouping suggests that the

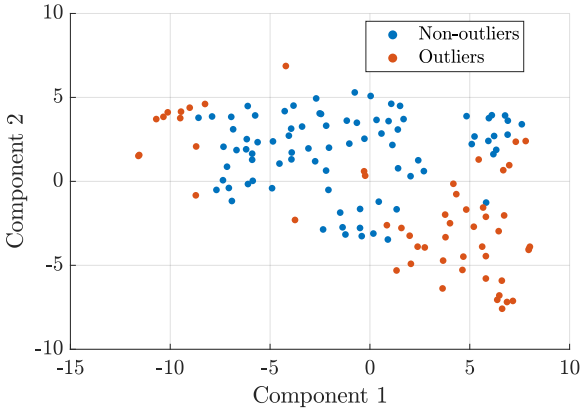


FIGURE 7: t-SNE visualization using K-Means and considering the 80 best-ranked features

"usable signals" structure is likely an inherent characteristic of our data, rather than an artifact of a specific clustering method.

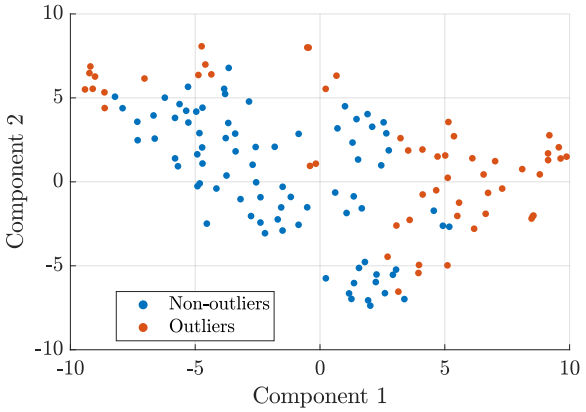


FIGURE 8: t-SNE visualization using Agglomerative Clustering and considering the 80 best-ranked features

Though both methods discarded usable signals and considered some usable signals unusable, Figure 9a shows a signal that was classified as usable but isn't. In Figure 9b, a usable signal was discarded by the outlier detection method. The difficulty in such cases is that visual inspection can only assess it as a misclassification. Using this procedure, we can make fine adjustments to the final result to achieve something closer to the ideal.

C. COMPARISON BETWEEN USABLE AND HEALTHY SIGNALS

This section demonstrates the clinical utility of the quality-controlled dataset by comparing physiological features between cleaned lesion signals (n=82) and healthy controls (n=24). These differences validate the cleaned dataset's suitability for downstream physiological analysis.

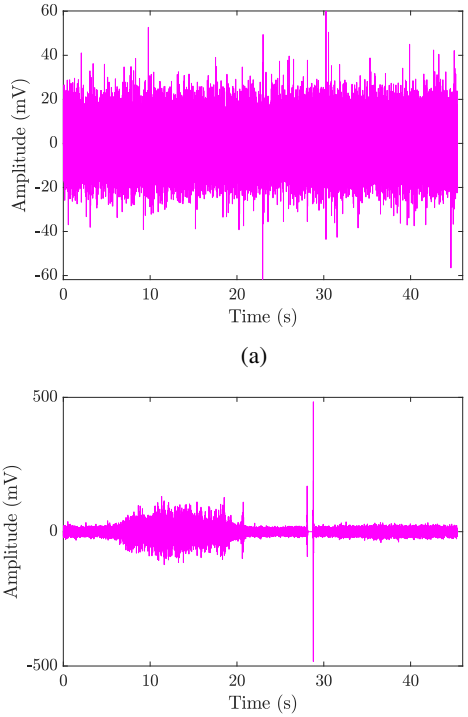


FIGURE 9: Misclassified signals. In (a), an unusable signal classified as usable is presented. In (b), a usable signal is presented classified as unusable.

At first, features were extracted from the signals of the control group and stored in a data frame. After that, a label column was inserted in both feature data frames with two values: "Lesion" and "Healthy". Next, these data frames were concatenated into one and used to visualize the data separability, as shown in Figure 10.

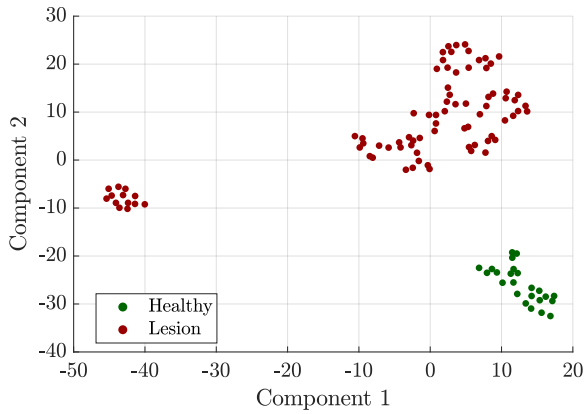


FIGURE 10: t-SNE visualization of usable and healthy signals

As observed, control signals form a well-defined cluster, while lesion signals are more dispersed. This behavior can be

attributed to a significant difference in average amplitudes. To ensure physiological interpretability, raw digital units were converted to voltage (V_{mV}) using the standard ADC transfer function: $V_{\text{mV}} = \frac{\text{ADC}_{\text{raw}}}{4095} \times 3300$. Following this conversion, the lesion group typically exhibited average amplitudes below 1,050 mV, whereas the control group exceeded 3,220 mV. These differences are often associated with muscular underuse or atrophy in individuals with neurological injuries. The clear separation between the two groups suggests a separability boundary within the feature space, which could form the basis for a classification model—though classification is not the focus of this study.

We also analyzed the average high-frequency energy levels. As shown in Figure 11, most signals in the lesion group exhibit high-frequency energy values near zero, with only a few outliers. This suggests a lack of high-frequency components in signals from individuals with neurological injuries. In contrast, signals from the control group show significantly higher high-frequency energy, reinforcing that this feature is strongly associated with muscle activity in healthy individuals. The clear separation between groups based on this feature also highlights its potential for use in simple decision-making systems.

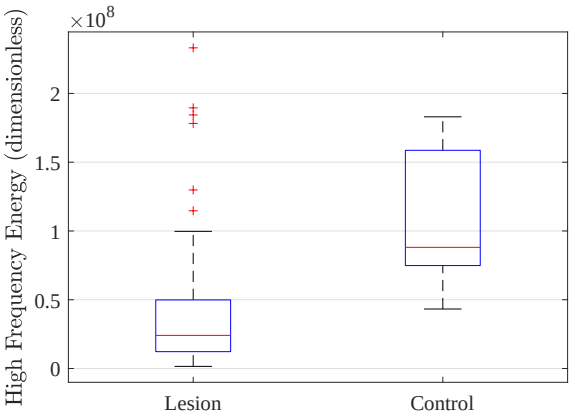


FIGURE 11: Comparison of high-frequency energy between Injury and Control groups.

Given the clear separation in high-frequency energy levels between the groups, this feature is a promising candidate for lightweight, threshold-based analysis. A rapid sweep through potential thresholds for energy values could validate the robustness of this separation. Moreover, such an approach would be computationally inexpensive and well-suited for real-time implementation in low-cost embedded systems. This insight highlights the potential to build fast and efficient decision logic without complex models—a key advantage when working under hardware or latency constraints.

To visualize the frequency spectrum of the average signals from the datasets, we calculated the median frequency, the average frequency, and the frequency ratio, as shown in Figure 12.

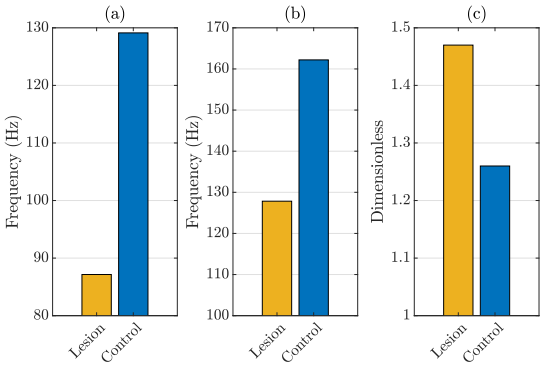


FIGURE 12: Frequency features for the comparison. In (a) is presented the median frequency. The average frequency is presented in (b). In (c), the frequency ratio is presented.

Furthermore, for the healthy signals, the median frequency is 1.42 times higher, the average frequency is 1.24 times higher, and the frequency ratio is lower, at 0.87 for the signals of people with injuries. These results suggest a higher concentration of energy in specific frequency bands in the signals of individuals with lesions, particularly in the low-frequency components, which are in accordance with three of the patients in the study presented by [35].

To verify these possible differences, the Fourier Transform with a Hamming window was performed on all signals, and three frequency metrics were extracted (median, mean, and the relationship between them). The statistical tests applied were the t-test (assuming equal variances), the Welch t-test (does not assume equal variances), the Mann-Whitney test (nonparametric), the Kolmogorov-Smirnov test (difference in distribution), normality tests (Kolmogorov-Smirnov), and calculation of effect size (Cohen's d).

To verify these differences statistically, we first assessed the normality of the data distributions using the Kolmogorov-Smirnov test. The results indicated that the Mean Frequency and Frequency Ratio followed a normal distribution for both groups ($p > 0.05$). However, the Median Frequency for the injury group did not follow a normal distribution ($p = 0.0047$).

Consequently, we applied the Welch's t-test (which does not assume equal variances) for the normally distributed features and the non-parametric Mann-Whitney U test for the Median Frequency. The effect size was calculated using Cohen's d to quantify the magnitude of the differences.

Table 5 summarizes the statistical comparison. All three spectral features showed highly significant differences between the Injury and Control groups ($p < 0.001$). Specifically, the Median Frequency presented a Cohen's d of **0.72**, indicating a medium-to-large effect size. This substantial effect size confirms that the spectral shift observed in the injured group (lower frequencies) is not merely a statistical artifact but reflects a distinct physiological alteration in motor unit recruitment and conduction velocity compared to healthy controls.

TABLE 5: Statistical comparison of spectral features between Injury and Healthy groups.

Feature	Normality	Test Used	p-value	Cohen's d
Median Freq.	Non-Normal	Mann-Whitney	< 0.001	0.66
Mean Freq.	Normal	Welch's t-test	< 0.001	0.81
Freq. Ratio	Normal	Welch's t-test	< 0.001	0.71

*Normality tested via Kolmogorov-Smirnov. Significant at $p < 0.05$.

According to Figure 13a, the Injury group presented a significantly lower median frequency, suggesting possible muscle fatigue or reduced recruitment of high-frequency fibers. The average frequency, Figure 13b, was even more sensitive to the difference between groups, with a large effect size, reinforcing the change in the spectral contents of the EMG signals in the Injury group. The frequency ratio, Figure 13c, indicates higher asymmetry in the frequency spectrum of the EMG signals in the Injury group, possibly reflecting neuromuscular changes related to the injury or fatigue.

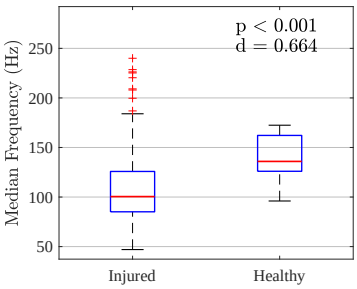
As shown in Figure 14, the frequency spectrum of the average signals for the two groups is quite different. The individuals with injuries have a spectrum with predominantly low-frequency components and very few high-frequency components. In contrast, the spectrum for the healthy control group shows components present throughout the ideal EMG signal range (10 - 450 Hz).

D. COMPARISON WITH RELATED WORKS

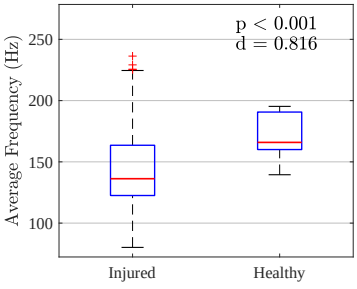
In addition, we bring other datasets related to people with neurological conditions. More specifically, the individuals whose sEMG data were acquired had a diagnosis of stroke. We gathered the main information from each dataset and studies that used them in Table 6. Each work analyzed different movements. The acquisition systems used in those studies were Trigno from Delsys [36]–[38] and the Myo Armband [39].

Unlike these existing datasets, which often provide pre-curated signals from strictly controlled laboratory settings, our work addresses the real-world variability of clinical acquisition. With a noise prevalence of nearly 40% identified in hospital settings (as shown in our Ground Truth analysis), our dataset provides a realistic benchmark for developing robust quality control algorithms. Furthermore, by including individuals with Cerebral Palsy and utilizing open-source hardware (ESP32), we extend the applicability of sEMG analysis beyond post-stroke recovery and high-cost proprietary systems.

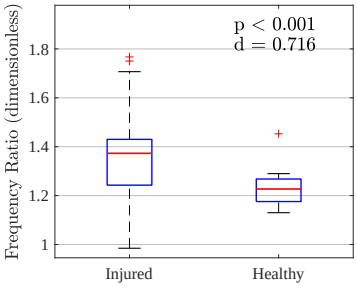
Although the movements differed for each study, the wrist extension and hand-reaching movements coincide with our present study. Furthermore, the number of volunteers is also different for each dataset. The present study includes, in addition to post-stroke individuals, individuals with cerebral palsy, providing electromyography data from individuals with a different neurological condition.



(a)



(b)



(c)

FIGURE 13: Distribution of the features used in the comparison. In (a), the distribution of the median frequency is presented. In (b), the average frequency distribution is presented. The frequency ratio is presented in (c).

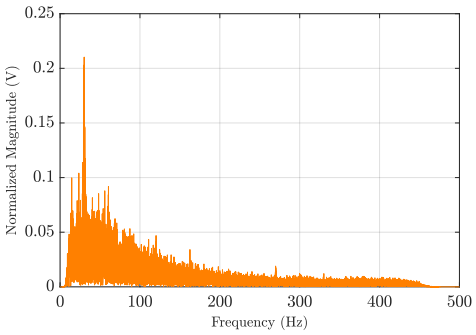
IV. LIMITATIONS

The dataset comprises signals from only 5 injured participants, with highly variable trial numbers per patient (Table 1: P1=58, P2=5, P3=46, P4=30, P5=6 trials). This imbalance and limited sample size restrict statistical power and generalizability to broader neurological populations. Future studies should include larger, more balanced cohorts across multiple injury types and severities.

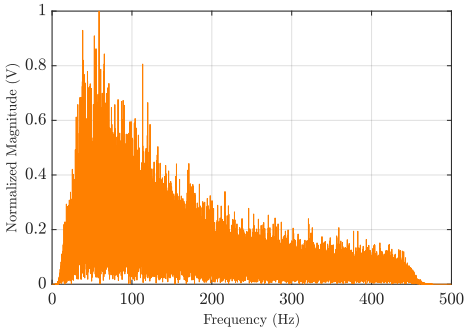
The quality assessment pipeline was developed and validated exclusively on data acquired with custom hardware (ESP32 plus AD8232 circuit). Performance on commercial EMG systems (e.g., Trigno Delsys, Myo Armband) or different electrode configurations remains untested. Electrode-skin impedance, gain settings, and sampling rates may affect feature extraction and clustering results, requiring cross-platform validation.

TABLE 6: Summary of EMG datasets in populations with neurological injury.

Dataset	Problem	Participants	Gestures	Reference	Technique	Acquisition System
MUSED-I	Stroke	2	Wrist flexion/extension, hand closing, wrist radial deviation, wrist ulnar deviation, and resting position	[39]	Feature extraction, classification, and pattern recognition	Myo Armband
EMG data of stroke and healthy subjects when performing hand-to-nose movement	Stroke	20	Hand-to-nose movement	[36]	Feature extraction, feature selection and regression	Trigno
EMG datasets from hand-reaching movements for multiple directions of healthy and post-stroke individuals	Post-Stroke	13	Maximum Voluntary Contraction (MVC), hand-reaching movements for different directions	[37], [38]	[37]Feature extraction, muscle synergy extraction, clustering coefficient, global efficiency, and betweenness-centrality. [38]Data segmentation and muscle synergies identification.	Trigno
sEMG data from people with cerebral palsy and post-stroke	Cerebral Palsy and Post-stroke	5	Elbow extension, wrist extension, and hand-reaching	[19]	Feature extraction, feature selection, ablation, clustering, and outlier detection	Developed hardware



(a)



(b)

FIGURE 14: Frequency spectra. In (a), the frequency spectrum of the injured group is presented. In (b), the frequency spectrum of the healthy group is presented.

Although objective criteria were established post-hoc (SNR, saturation thresholds), the initial "usable/unusable" ground truth relied on expert visual inspection, introducing potential subjectivity. Inter-rater reliability was not formally assessed, and borderline cases may exist. Automated methods achieved excellent agreement (ARI = 0.85), but residual bias from human annotation persists.

Data collection occurred at one institution (Pequeno Cotolengo Health Complex, Brazil) from participants with post-stroke and cerebral palsy. Demographic homogeneity (age, ethnicity, lesion chronicity) and environmental factors (hospital setting) limit external validity. Multi-center studies across diverse populations and clinical environments are needed.

A methodological limitation lies in the definition of the Ground Truth itself. The classification of signals as 'Usable' or 'Unusable' relied on fixed empirical thresholds (e.g., SNR < 6 dB, saturation $> 1.0\%$). As highlighted by recent reviews [20], the objective identification of artifacts in sEMG remains an open challenge with no universally standardized consensus. Consequently, strict thresholds may inadvertently categorize marginally noisy but physiologically valid signals as unusable—a limitation observed in our False Positive analysis. This finding reinforces the relevance of the proposed unsupervised approach, which groups signals based on intrinsic data structure rather than rigid, potentially arbitrary cutoff values.

The method processes complete 45s signals rather than streaming data. Real-time implementation during acquisition (crucial for immediate re-collection decisions) requires additional latency and computational validation.

These limitations inform our future research directions: (i) multi-center validation comparing the custom ESP32-

based system against gold-standard commercial hardware; (ii) expansion to larger cohorts to assess clinical inter-rater reliability; (iii) optimization of the unsupervised pipeline for real-time streaming on embedded devices; and (iv) the development of adaptive, consensus-based quality metrics to overcome the reliance on fixed empirical thresholds, addressing the current lack of standardization in the field.

V. CONCLUSIONS

This study demonstrated that an unsupervised machine learning pipeline can effectively automate the quality assessment of sEMG signals from individuals with neurological injuries, addressing the limitations of subjective visual inspection. By combining clustering with anomaly detection, the proposed method successfully distinguished usable physiological data from acquisition artifacts without requiring labeled training data.

Crucially, the physiological validation confirmed that the retained signals preserve the distinct spectral and amplitude characteristics associated with neuromuscular pathology, such as muscular atrophy and lower firing frequencies. This indicates that the automated cleaning process maintains clinical validity. The results also highlight the complexity of defining a binary ground truth for pathological sEMG, suggesting that unsupervised approaches may be more robust than rigid thresholds for capturing the variability of injured muscle signals. Future work will focus on optimizing the pipeline for real-time processing on embedded devices and validating the approach against gold-standard commercial systems in multi-center studies.

REFERENCES

[1] J. D. Steinmetz, et al., "Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021," *The Lancet Neurology*, vol. 23, no. 4, p. 344–381, Apr. 2024. [Online]. Available: http://dx.doi.org/10.1016/S1474-4422(24)00038-3

[2] Ministério dos Direitos Humanos e da Cidadania. (2023) O brasil tem cerca de 18,9 milhões de pessoas com deficiência. [Online]. Available: https://www.gov.br/mdh/pt-br/navegue-por-temas/pessoa-com-deficiencia/estatisticas

[3] M. A. Delph, S. A. Fischer, P. W. Gauthier, C. H. M. Luna, E. A. Clancy, and G. S. Fischer, "A soft robotic exomusculature glove with integrated semg sensing for hand rehabilitation," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, 2013, pp. 1–7.

[4] S. Huang, S. Cai, G. Li, Y. Chen, K. Ma, and L. Xie, "semg-based detection of compensation caused by fatigue during rehabilitation therapy: A pilot study," *IEEE Access*, vol. 7, pp. 127 055–127 065, 2019.

[5] L. McManus, G. De Vito, and M. M. Lowery, "Analysis and biophysics of surface emg for physiotherapists and kinesiologists: Toward a common language with rehabilitation engineers," *Frontiers in Neurology*, vol. 11, Oct. 2020. [Online]. Available: http://dx.doi.org/10.3389/fneur.2020.576729

[6] T. Song, Z. Yan, S. Guo, Y. Li, X. Li, and F. Xi, "Review of semg for robot control: Techniques and applications," *Applied Sciences*, vol. 13, no. 17, p. 9546, Aug. 2023. [Online]. Available: http://dx.doi.org/10.3390/app13179546

[7] D. Yadav and K. Veer, "Recent trends and challenges of surface electromyography in prosthetic applications," *Biomedical Engineering Letters*, vol. 13, no. 3, p. 353–373, Apr. 2023. [Online]. Available: http://dx.doi.org/10.1007/s13534-023-00281-z

[8] G. Coraggio, M. Cera, M. Cirelli, and P. P. Valentini, "Review and comparison of linear algorithms to quantify muscle fatigue based on semg signals," *Ergonomics*, vol. 67, no. 11, p. 1729–1747, May 2024. [Online]. Available: http://dx.doi.org/10.1080/00140139.2024.2349962

[9] S. Wang, H. Tang, B. Wang, and J. Mo, "A novel approach to detecting muscle fatigue based on semg by using neural architecture search framework," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 4932–4943, 2023.

[10] Y. Zhou, C. Chen, M. Cheng, Y. Alshahrani, S. Franovic, E. Lau, G. Xu, G. Ni, J. M. Cavanaugh, S. Muh, and S. Lemos, "Comparison of machine learning methods in semg signal processing for shoulder motion recognition," *Biomedical Signal Processing and Control*, vol. 68, p. 102577, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.bspc.2021.102577

[11] A. Tigrini, A. H. Al-Timemy, F. Verdini, S. Fioretti, M. Morettini, L. Burattini, and A. Mengarelli, "Decoding transient semg data for intent motion recognition in transhumeral amputees," *Biomedical Signal Processing and Control*, vol. 85, p. 104936, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.bspc.2023.104936

[12] D. P. Campos, J. J. A. Mendes Junior, P. B. Junior, A. E. Lazzaretti, L. G. Sartori, and E. Krueger, "Non-invasive muscle-machine interface open source project: wearable hand myoelectrical orthosis (mes-fes)," *Assistive Technology*, p. 1–10, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1080/10400435.2024.2382857

[13] D. B. Ribeiro, L. G. Sartori, M. V. G. Méndez, R. B. d. Souza, D. P. Campos, P. B. Júnior, J. J. A. M. Junior, and E. Krueger, "Mes-fes interface enhances quadriceps muscle response in sitting position in incomplete spinal cord injury: Pilot study," *Prosthesis*, vol. 6, no. 3, p. 643–656, Jun. 2024. [Online]. Available: http://dx.doi.org/10.3390/prosthesis6030045

[14] C. Marquez-Chin and M. R. Popovic, "Functional electrical stimulation therapy for restoration of motor function after spinal cord injury and stroke: a review," *BioMedical Engineering OnLine*, vol. 19, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218873586

[15] M. Boyer, L. Bouyer, J.-S. Roy, and A. Campeau-Lecours, "Reducing noise, artifacts and interference in single-channel emg signals: A review," *Sensors*, vol. 23, no. 6, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/6/2927

[16] I. Campanini, C. Disselhorst-Klug, W. Z. Rymer, and R. Merletti, "Surface emg in clinical assessment and neurorehabilitation: Barriers limiting its use," *Frontiers in Neurology*, vol. 11, 9 2020.

[17] J. Mendes Júnior, D. Campos, L. Biassio, P. Passos, P. Júnior, A. Lazzaretti, and E. Krueger, "Ad8232 to biopotentials sensors: Open source project and benchmark," *Electronics*, vol. 12, p. 833, 02 2023.

[18] A. T. P. Inafuco, P. Machoski, D. P. Campos, S. F. Pichorim, and J. J. A. Mendes Junior, "Mot: A low-latency, multichannel wireless surface electromyography acquisition system based on the ad8232 front-end," *Sensors*, vol. 25, no. 12, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/12/3600

[19] M. V. G. Méndez, "Evaluation of an semg-fes neuro-orthesis system: a case study in neurological patients," Master's thesis, Federal Technological University of Paraná, 2025, http://repositorio.utfpr.edu.br/jspui/handle/1/36141.

[20] M. Ait Yous, S. Agounad, and S. Elbaz, "Detection, identification and removing of artifacts from semg signals: Current studies and future challenges," *Computers in Biology and Medicine*, vol. 186, p. 109651, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482525000010

[21] P. V. Komi and P. Tesch, "Emg frequency spectrum, muscle structure, and fatigue during dynamic contractions in man," *European Journal of Applied Physiology and Occupational Physiology*, vol. 42, no. 1, p. 41–50, Sep. 1979. [Online]. Available: http://dx.doi.org/10.1007/BF00421103

[22] C. J. De Luca, "The use of surface electromyography in biomechanics," *Journal of Applied Biomechanics*, vol. 13, no. 2, p. 135–163, May 1997. [Online]. Available: http://dx.doi.org/10.1123/jab.13.2.135

[23] R. Merletti and P. Di Torino, "Standards for reporting emg data," *J Electromyogr Kinesiol*, vol. 9, no. 1, pp. 3–4, 1999.

[24] H. Tankisi, D. Burke, L. Cui, M. de Carvalho, S. Kuwabara, S. D. Nandedkar, S. Rutkove, E. Stålberg, M. J. van Putten, and A. Fuglsang-Frederiksen, "Standards of instrumentation of emg," *Clinical Neurophysiology*, vol. 131, no. 1, p. 243–258, Jan. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.clinph.2019.07.025

[25] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE transactions on biomedical engineering*, vol. 40, no. 1, pp. 82–94, 1993.

[26] Y. Niu, W. Chen, H. Zeng, Z. Gan, and B. Xiong, "Optimizing semg gesture recognition: Leveraging channel selection and feature

compression for improved accuracy and computational efficiency,'' *Applied Sciences*, vol. 14, no. 8, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/8/3389

[27] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, ''Tsfel: Time series feature extraction library,'' *SoftwareX*, vol. 11, p. 100456, 2020.

[28] L. van der Maaten and G. Hinton, ''Visualizing data using t-sne,'' *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[29] F. T. Liu, K. M. Ting, and Z.-H. Zhou, ''Isolation forest,'' in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[30] G. Balbinot, G. Li, M. J. Wiest, M. Pakosh, J. C. Furlan, S. Kalsi-Ryan, and J. Zariffa, ''Properties of the surface electromyogram following traumatic spinal cord injury: a scoping review,'' *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, Jun. 2021. [Online]. Available: http://dx.doi.org/10.1186/s12984-021-00888-2

[31] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, ''Time-series clustering – a decade review,'' *Information Systems*, vol. 53, pp. 16–38, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437915000733

[32] F. Murtagh and P. Contreras, ''Algorithms for hierarchical clustering: an overview,'' *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, p. 86–97, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1002/widm.53

[33] P. A. McNulty, G. Lin, and C. G. Doust, ''Single motor unit firing rate after stroke is higher on the less-affected side during stable low-level voluntary contractions,'' *Frontiers in Human Neuroscience*, vol. 8, Jul. 2014. [Online]. Available: http://dx.doi.org/10.3389/fnhum.2014.00518

[34] X. Hu, A. K. Suresh, W. Z. Rymer, and N. L. Suresh, ''Assessing altered motor unit recruitment patterns in paretic muscles of stroke survivors using surface electromyography,'' *Journal of Neural Engineering*, vol. 12, no. 6, p. 066001, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1088/1741-2560/12/6/066001

[35] S. Angelova, S. Ribagin, R. Raikova, and I. Veneva, ''Power frequency spectrum analysis of surface emg signals of upper limb muscles during elbow flexion – a comparison between healthy subjects and stroke survivors,'' *Journal of Electromyography and Kinesiology*, vol. 38, pp. 7–16, 2018, neuromechanics of fine hand-motor tasks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641117301207

[36] K. Zhao, H. Wen, Z. Zhang, C. He, and J. Wu, ''Fractal characteristics-based motor dyskinesia assessment,'' *Biomedical Signal Processing and Control*, vol. 68, p. 102707, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421003049

[37] K. Zhao, Y. Feng, L. Li, Y. Zhou, Z. Zhang, and J. Li, ''Muscle synergies and muscle networks in multiple frequency components in post-stroke patients,'' *Biomedical Signal Processing and Control*, vol. 95, p. 106417, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809424004750

[38] S. Israely, G. Leisman, C. Machluf, T. Shnitzer, and E. Carmeli, ''Direction modulation of muscle synergies in a hand-reaching task,'' *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 12, pp. 2427–2440, 2017.

[39] A. Winursito, F. Arifin, M. Muslikhin, H. Artanto, and F. H. Caryn, ''Performance analysis of emg signal classification methods for hand gesture recognition in stroke rehabilitation,'' *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 8, no. 2, p. 264–274, Aug. 2024. [Online]. Available: http://dx.doi.org/10.21831/elinvo.v8i2.76811

ACKNOWLEDGMENT

This study was financed by the National Council for Scientific and Technological Development (CNPq - Brazil).

JOÃO PEDRO M. LOURENÇÃO received the B.S. degree in computer engineering from Federal Technological University - Paraná (UTFPR), Apucarana, Paraná, Brazil, in 2024. He is currently pursuing a Master´s Degree in biomedical engineering at the Postgraduate Program in Electrical Engineering and Industrial Informatics (UTFPR), Curitiba, Paraná, Brazil.

HEITOR S. LOPES is a Full Professor at Federal Technological University - Paraná. He received the B.S. degree in electronic engineering from Federal Technological University - Paraná (UTFPR) in 1984, received a Master´s degree in Science in Biomedical Engineering in 1990, and received a Ph.D. degree in Engineering in 1996.

MARÍA VERÓNICA G. MÉNDEZ is a physiotherapist who graduated from the State University of Londrina (UEL), as a participant in the Brazil-Paraguay Undergraduate Student Exchange Program. Specialist in Neurofunctional Physiotherapy (Adult) from UEL. Elective internship in the Multiprofessional Residency in Health Promotion and Care in Hospital Attention, area of concentration adult and elderly health, at the University of São Paulo (USP). Training in the basic PediaSuit protocol and intensive therapy with the Skills Cage. Master's student in Electrical Engineering and Industrial Informatics, area of concentration Biomedical Engineering, at the Federal Technological University of Paraná (UTFPR) in Curitiba.

CRISTIAN VEGGIAN MATIAS earned his B.S. in Computer Engineering from UTFPR Apucarana, Brazil, graduating in 2024. Currently, he pursues a Master's degree in Electrical and Computer Engineering (CPGEI) at UTFPR Curitiba, Brazil. His work centers on machine learning for sEMG, ECG, and EEG person identification in robotic neurorehabilitation, along with biosignal analysis and Python-based development.

DANIEL P. CAMPOS is a Professor at Federal Technological University - Paraná. He received a B.S. degree in Electrical Engineering (2014), a Master's degree in Electrical Engineering from the Electrical Engineering Graduate Program (PPGEE) (2016), PhD in Biomedical Engineering from the Electrical Engineering and Industrial Informatics Graduate Program (CPGEI) (2019), all from the Federal University of Technology - Paraná (UTFPR). Completed a postdoctoral fellowship in Neural Engineering and Rehabilitation at the State University of Londrina (UEL).

JOSÉ JAIR A. MENDES JÚNIOR received the Automation Technologist, M.Sc., and D.Sc. degrees in electrical engineering from the Federal University of Technology—Paraná, Curitiba, in 2015, 2016, and 2020, respectively, and the B.Sc. degree in electrical engineering from the Higher Education Center, Campos Gerais, in 2020. He is currently a Professor in DAELN, Federal University of Technology — Paraná at Curitiba, Brazil. His research interests include biomedical instrumentation, signal processing, machine learning, and industrial automation.