Performance Analysis of Machine Learning Models for Hand Gesture Recognition from Electromyographic Signals

Lucas Lemos Cerqueira de Freitas[1], **Artur Brederodes da Costa Neto**[1], **Gabriel Lucas Bento Germano**[1], **Maria Fernanda Herculano Machado da Silva**[2], **Rodrigo Santos da Silva**[3], **Thiago Damasceno Cordeiro**[1]

[1]Computing Institute – Federal University of Alagoas (UFAL)

[2]Department of Electrical Engineering – Federal University of Minas Gerais (UFMG)

[3]Department of Computer Science – Federal University of Minas Gerais (UFMG)

thiago@ic.ufal.br

Abstract. *Amputation of the upper extremity significantly affects human autonomy by compromising motor functionality, limiting the execution of activities of daily living, and reducing the individual's capacity for independent interaction with the environment. In 2017, more than 57 million people were living with some form of traumatic amputation, approximately 30% involving the upper limbs. In this context, prostheses controlled by surface electromyographic (sEMG) signals are a promising alternative for restoring motor function, using residual muscles' electrical activity as a control source. However, automatic gesture recognition from sEMG signals remains challenging, due to physiological variability, noise, and the complexity of neuromuscular patterns. This work presents a systematic and reproducible comparative analysis of four supervised classifiers — KNN, Random Forest, SVM, and Multilayer Perceptron — applied to the NinaPro DB5 dataset, with focus on the joint effect of window size and temporal stride on classification performance, a combination rarely explored in the literature. Time-domain feature extraction and sliding-window segmentation were applied to the models. The KNN and RF models achieved the best results, with accuracies exceeding 99% in certain configurations, reinforcing the potential of supervised methods for developing more accessible and responsive robotic prostheses and human-machine interfaces.*

1. Introduction

Upper limb amputation imposes significant functional, emotional, and social limitations, affecting the autonomy and quality of life of millions. In 2017, more than 57 million people worldwide were living with traumatic limb amputations, approximately 30% involving the upper limbs [McDonald et al. 2021]. The loss of hand function is particularly debilitating, given its central role in both basic daily tasks — such as eating and dressing—and complex activities — such as typing or manipulating delicate objects. In this context, upper limb prostheses have emerged as an important technological solution with the potential to partially restore lost functionality. Among the available approaches, prostheses controlled by surface electromyographic (sEMG) signals stand out for their ability to capture the electrical activity of residual muscles and translate it into commands for artificial movements. Recent advances in machine learning have enabled more precise and

intuitive control systems, reducing user cognitive effort and increasing the naturalness of prosthetic gestures [Corbett et al. 2011, Kadavath et al. 2024].

Electromyography (EMG) records and analyzes the electrical activity of skeletal muscles during contraction, enabling non-invasive investigation of neuromuscular control [Mills 2005, Garcia and Vieira 2011]. This activity consists of action potentials—rapid voltage changes in muscle fiber membranes—generated when impulses from the central nervous system activate ion channels [De Luca 2002]. These potentials propagate along muscle fibers toward tendinous regions, producing contraction patterns that reflect neuromuscular recruitment.

EMG signals can be obtained using intramuscular or surface electrodes. Intramuscular electrodes allow analysis of individual motor units but are invasive and typically restricted to clinical contexts [Mills 2005]. Surface EMG (sEMG), in contrast, records the overall electrical activity of muscle groups via electrodes placed on the skin, providing a safe, non-invasive approach [Garcia and Vieira 2011, De Luca 2002].

As shown in Figure 1, sEMG acquisition involves recording motor unit activity via skin electrodes, followed by amplification, filtering, and digitization for analysis. These properties make sEMG suitable for real-time applications such as gesture recognition in robotic prostheses and assistive human–machine interfaces [Garcia and Vieira 2011, Kadavath et al. 2024, Phinyomark et al. 2012].

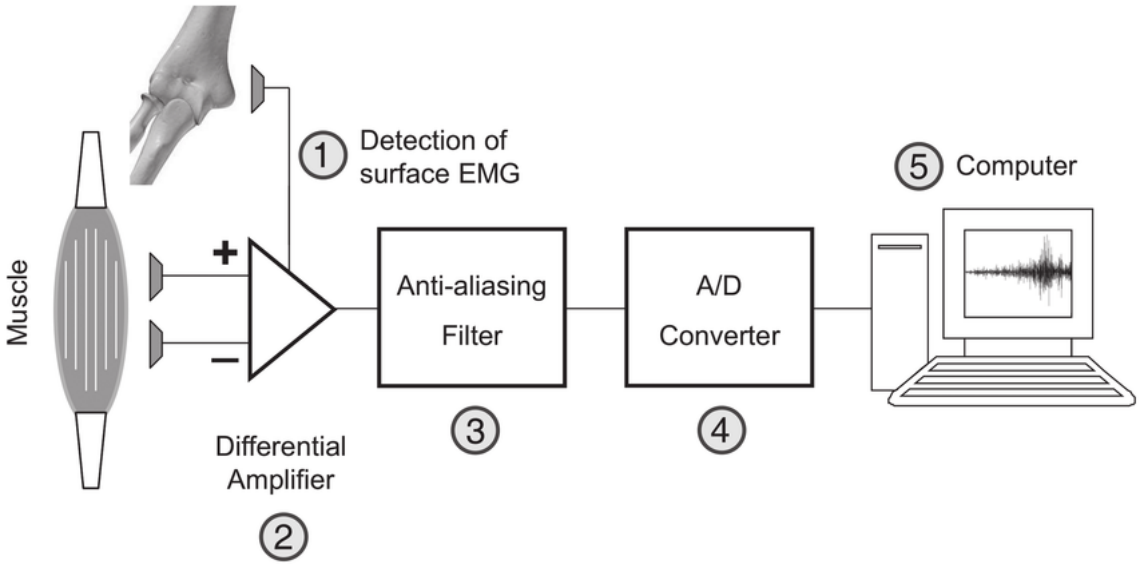

Figure 1. Simplified diagram of the sEMG signal acquisition process.

However, efficient and accurate control of sEMG-driven prostheses remains a challenge. Inter- and intra-user signal variability, the presence of physiological and environmental noise, and the complexity of muscular patterns make gesture recognition an active and continuously evolving research area. One of the main bottlenecks lies in the signal classification stage, which must accurately distinguish different hand gestures even when performed at varying intensities or with subtle variations. Prior work has explored hybrid machine-learning and deep-learning approaches to improve gesture recognition for amputee users, highlighting the importance of appropriately selected temporal windows and preprocessing parameters for model generalization [Gopal et al. 2022]. Other studies

have shown that classical models, such as Random Forest, can still outperform neural networks in certain contexts, achieving accuracies above 99% in multi-gesture classification using the Myo Armband sensor [Kadavath et al. 2024]. Nevertheless, systematic comparisons of multiple supervised classifiers across varying temporal window configurations on public datasets such as NinaPro remain limited, as most studies focus on specific aspects, such as deep architectures or amputee populations, without broadly discussing the impact of algorithm choice and windowing parameters on performance.

This work addresses that gap by providing a systematic and reproducible comparative analysis of four supervised classifiers — K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) — applied to the publicly available NinaPro DB5 dataset. The main contribution of this study is the joint and controlled evaluation of the effect of window size and temporal stride on classification performance across multiple algorithms, a combination rarely explored in the literature, where analyses typically focus either on a single classifier or on a fixed segmentation configuration. By systematically varying five window sizes (50-400 samples) and multiple stride lengths, along with time-domain feature extraction and hyperparameter optimization via GridSearchCV, this work provides a comprehensive view of the robustness, efficiency, and sensitivity of each method. The results contribute to a deeper understanding of the factors that govern the accuracy and practical viability of myoelectric gesture recognition systems, providing evidence-based guidelines for the design of more accessible and responsive robotic prostheses.

2. Methodology

2.1. Database conversion and organization

The NinaPro database provides open-source data essential for the development and evaluation of pattern-recognition algorithms for robotic prosthetic control. In this work, the DB5 *dataset*, which consists exclusively of signals from healthy individuals, was used to model natural hand and wrist gestures in a non-invasive control scenario.

The acquisition of sEMG signals was conducted using two Thalmic Myo Armband devices, positioned around the dominant forearm of each participant, covering the flexor and extensor muscles. The skin-electrode interface was cleaned with isopropyl alcohol to reduce impedance. The setup comprised 16 active sEMG channels, with sensors operating at 200 Hz, 8-bit resolution, and built-in low-pass filtering (5–100 Hz).

During the experiments, the 10 participants performed various movements interleaved with rest periods (3 seconds) to avoid muscle fatigue. Aiming to reduce the complexity of the classification problem and focus on different types of activation, a subset of 7 gestures was selected, repeated six times per participant, with an average duration of 5 seconds each: Index finger flexion, Thumbs up, Extension of index and middle fingers, flexion of the others, Opening (abduction) of all fingers, Fingers flexed, forming a fist, Pointing index finger, and Medium wrap grasping of an object.

The original signals of each individual in the DB5 were distributed across three `.mat` files, corresponding to different exercise groups (A, B, and C). To unify this information into a manageable structure, a reading and concatenation process was implemented using the Python libraries *NumPy* (v1.26.4) and *Pandas* (v2.2.3). Adjustments to

the stimulus labels (*stimulus*) were applied to ensure the correct indexing of movements after unification.

2.2. Signal pre-processing

No additional digital filtering was applied to the raw signals, as the sensor hardware filtering provided an adequate signal-to-noise ratio. A descriptive exploratory analysis was conducted to evaluate the distribution of variables and data integrity (Figures 2 and 3).

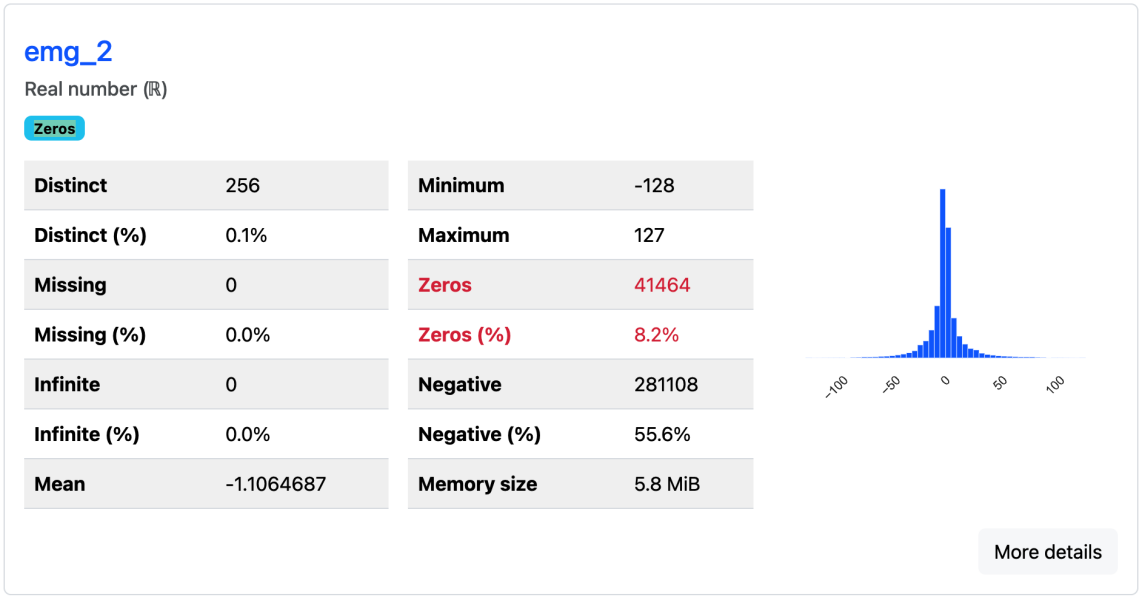

Figure 2. Descriptive statistics and distribution of the sEMG signal in channel 2.

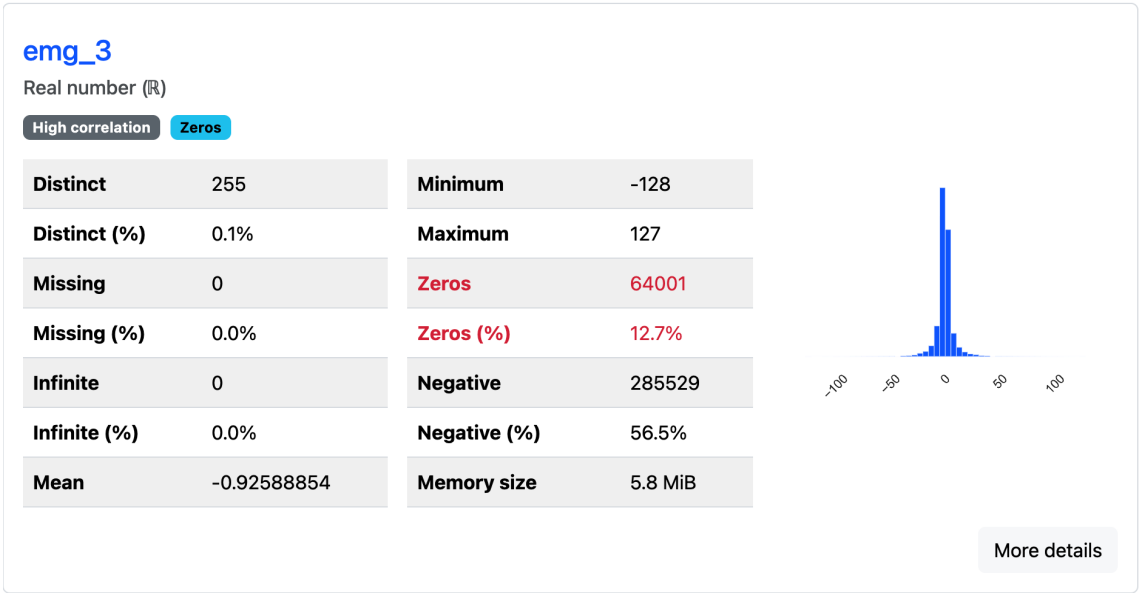

Figure 3. Descriptive statistics and distribution of the sEMG signal in channel 3.

To assess the need for dimensionality reduction due to redundancy, the Pearson correlation matrix for the 16 sEMG channels was computed (Table 1). A threshold of

Channel A	Channel B	Correlation Coeff. (r)
EMG10	EMG11	0.55
EMG3	EMG4	0.49
EMG11	EMG12	0.46
EMG2	EMG3	0.46
EMG4	EMG5	0.39
EMG9	EMG16	0.37
EMG1	EMG2	0.37
EMG6	EMG7	0.37
EMG7	EMG8	0.36
EMG5	EMG6	0.35

Table 1. Highest Pearson correlations between sEMG channels.

$r = 0.7$ was empirically established for channel removal. Since the highest observed correlation was 0.55 (between EMG10 and EMG11), all 16 channels were retained for feature extraction.

2.3. Windowing and feature extraction

Signal segmentation was performed using an overlapping sliding-window technique. To rigorously evaluate the impact of temporal resolution on classifier performance, five window widths (ranging from 50 to 400 samples) associated with different stride lengths were tested (Table 2). The stride was defined as a percentage of the window size; this approach ensures a constant overlap ratio and maintains a proportional sample density for training, promoting a fair comparison between configurations. From each temporal segment, a feature vector was extracted, composed of metrics widely validated in the myoelectric control literature [Oskoei and Hu 2007]: *Root Mean Square* (RMS), *Zero Crossing* (ZC), *Variance* (VAR), *Mean Absolute Value* (MAV), *Slope Sign Change* (SCC), and *Waveform Length* (WL).

Window Size (samples)	Duration (ms)	Stride length (%)
50	250	10%, 20%, 50%
100	500	10%, 25%
200	1000	5%, 10%, 25%
300	1500	5%, 10%, 25%
400	2000	5%, 10%, 25%

Table 2. Tested configurations for signal windowing.

3. Classification

For gesture recognition, the performance of different machine learning models trained on the feature vectors was evaluated. The dataset was partitioned using the stratified hold-out technique, with 80% of the samples for training and 20% for testing.

The *Dummy Classifier* was adopted as a baseline. Subsequently, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) models were applied, selected for their performance on the NinaPro database [Atzori et al. 2012].

Hyperparameter optimization was performed using an exhaustive grid search with cross-validation (*GridSearchCV*), following the search space detailed in Table 3. The classical models were implemented using *scikit-learn* (v1.5.0), while the MLP neural network was developed in *TensorFlow* (v2.17.0), both in a *Python* (v3.12.2) environment. The general pipeline adopted in this work is illustrated in Figure 4.

Model	Hyperparameter search space
MLP	`batch_size`: [32, 64, 128]; `epochs`: [50, 100]; `learning_rate`: [0.001, 0.005]
KNN	`n_neighbors`: [3, 5, 7, 9, 11]
RF	`n_estimators`: [100, 200]; `max_depth`: [None, 10, 20]; `min_samples_split`: [2, 5]; `min_samples_leaf`: [1, 2]
SVM	`kernel`: [linear, rbf]; `C`: [0.1, 1.0, 10.0]; `gamma`: [scale, auto]

Table 3. Hyperparameters evaluated via *GridSearchCV*.

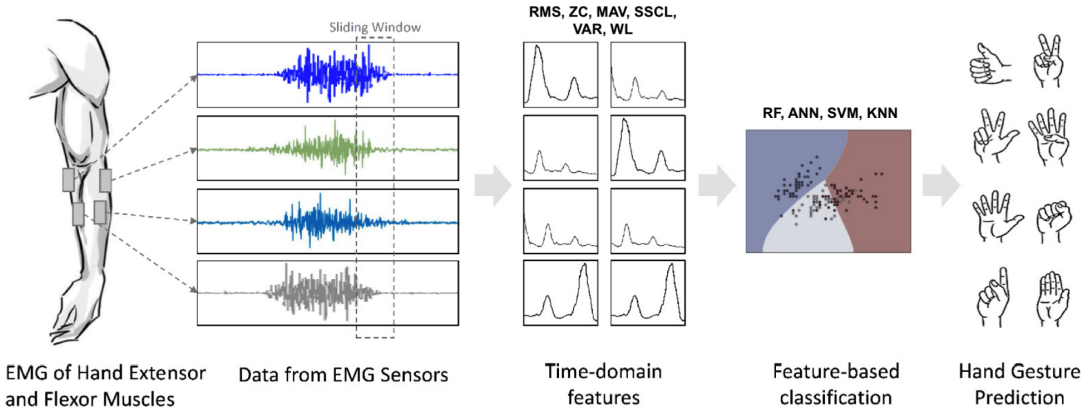

Figure 4. Flowchart of the feature extraction and classification process. Adapted from [Gopal et al. 2022].

4. Results and Discussion

All experiments were executed in an interactive Jupyter Notebook, running locally on a MacBook Air equipped with an Apple M3 processor, 8 GB of unified RAM, and 256 GB

of SSD storage.

4.1. Classifier performance

To facilitate interpretation and comparative analysis, the results were organized by the window size used in the signal segmentation. Thus, Tables 4, 5, 6, 7, and 8 present, respectively, the accuracies obtained by the classifiers for windows of 50, 100, 200, 300, and 400 samples. This approach allows for a clear observation of the impact of temporal granularity on the classifiers' performance.

Table 4. Classifier accuracy for a 50-sample window

Stride	Classifier				
	KNN	MLP	RF	SVM	Dummy
5	99.83%	89.78%	98.63%	92.64%	14.40%
10	97.05%	82.22%	94.61%	84.42%	14.47%
25	69.57%	70.41%	82.46%	69.57%	13.75%

Table 5. Classifier accuracy for a 100-sample window

Stride	Classifier				
	KNN	MLP	RF	SVM	Dummy
10	99.87%	93.42%	98.19%	95.94%	14.56%
25	94.53%	82.01%	90.13%	85.90%	13.90%

Table 6. Classifier accuracy for a 200-sample window

Stride	Classifier				
	KNN	MLP	RF	SVM	Dummy
10	99.84%	97.88%	99.48%	99.53%	14.72%
20	99.60%	96.24%	96.79%	97.70%	13.91%
50	92.78%	87.14%	88.62%	89.26%	13.65%

Table 7. Classifier accuracy for a 300-sample window

Stride	Classifier				
	KNN	MLP	RF	SVM	Dummy
15	99.81%	98.69%	99.08%	99.47%	14.82%
30	99.05%	97.86%	94.15%	97.68%	14.19%
75	94.14%	91.76%	86.79%	92.65%	12.47%

In addition to accuracy, the F1-score is an important complementary metric for evaluating performance in multiclass tasks, as it balances precision and recall. To avoid information overload and maintain focus on the most representative results, the F1-score

Table 8. Classifier accuracy for a 400-sample window

Stride	Classifier				
	KNN	MLP	RF	SVM	Dummy
20	99.49%	98.71%	98.42%	99.29%	14.12%
40	98.77%	96.83%	94.53%	97.98%	13.62%
100	89.31%	89.80%	85.45%	91.49%	13.76%

Table 9. Classifiers' F1-score for a 200-sample window and stride of 10

Gesture	Classifier			
	KNN	MLP	RF	SVM
Index finger flexed	99.89%	98.15%	99.89%	99.96%
Thumbs up	99.86%	98.45%	99.20%	99.37%
Index and middle fingers extended	99.89%	98.51%	99.29%	99.36%
Complete opening	99.75%	98.39%	99.22%	99.58%
Closed fist	99.69%	98.02%	99.32%	99.49%
Pointing index finger	99.87%	97.82%	99.66%	99.43%
Medium grasp	99.93%	98.43%	99.73%	99.56%

was calculated only for the configuration with a 200-sample window and a stride of 10, as shown in Table 9, since it presented the best overall performance, with high accuracy across all classifiers and good stability among them.

The obtained results reveal important patterns regarding the classifiers' performance as a function of the segmentation window size. In general, all supervised learning models outperformed the *Dummy* classifier by a wide margin, indicating that the sEMG signals contain relevant discriminative information for the gesture-recognition task.

It is observed that the KNN and RF classifiers achieved the best performance across practically all evaluated scenarios, with RF achieving accuracies above 98% for the 200- and 300-sample windows. This behavior is consistent with the literature, which reports the robustness of these models to noise and the high dimensionality of EMG data [Gopal et al. 2022, Kadavath et al. 2024]. Notably, Kadavath et al. [Kadavath et al. 2024] reported accuracies above 99% for RF applied to Myo Armband data — a result closely aligned with the findings of this work for the same classifier under optimal windowing configurations, despite differences in dataset and gesture set. Similarly, Gopal et al. [Gopal et al. 2022] observed that classical machine learning models remain competitive with deep learning approaches when appropriate preprocessing parameters are applied, a conclusion further supported by the present results.

Although the MLP model yielded satisfactory results, it was more sensitive to variations in segmentation parameters, especially in smaller windows. This can be attributed to the need for larger data volumes and the greater complexity of hyperparameter tuning, which is typical of artificial neural networks. Its performance, although slightly lower than the others, remained high, with F1-scores above 96% for all classes. This indicates that, despite its greater sensitivity to segmentation parameters, the model still adequately

generalized the muscular patterns associated with each gesture.

In contrast, the SVM's performance was competitive, especially for larger windows, where it achieved an accuracy close to 99%, highlighting its generalization capacity with wide margins in high-dimensional spaces. This result is consistent with previous findings by Oskoei and Hu [Oskoei and Hu 2007], who demonstrated the effectiveness of SVMs for myoelectric control tasks under similar high-dimensional feature conditions.

Another relevant point is the progressive degradation in the classifiers' performance as the window stride increases. This effect may be related to the lower temporal variability of adjacent windows, which reduces the diversity of the training data and leads to the overlap of redundant patterns. This observation corroborates previous studies that point out the importance of adjusting the overlap rate to avoid *overfitting* in classification tasks with temporal signals [Gopal et al. 2022, Tanaka et al. 2022].

The *Dummy* classifier, in turn, achieved an accuracy of approximately 14% across all windows, which is consistent with the random probability for seven classes ($\approx$14.3%). This result confirms that the learning models were capable of extracting relevant patterns from sEMG signals, substantially outperforming random classification.

Additionally, the analysis of the F1-score values as presented in Table 9 reinforces the robustness of the applied classifiers. The KNN, RF, and SVM classifiers achieved F1-scores above 99% across all gesture classes, demonstrating not only high overall accuracy but also a balance between precision and recall within each class. This result is particularly relevant in multiclass tasks, where accuracy can mask errors concentrated in one or more categories.

The F1-score analysis also enables the identification of potential asymmetries between classes. For example, the gestures "Pointing index finger" and "Thumbs up" exhibited the lowest F1-scores in the MLP (97.82% and 98.45%, respectively), which may indicate greater overlap between the myoelectric patterns of these classes or lower model robustness in identifying these specific variations. The absence of low F1-score values across classes reinforces the absence of neglect of minority classes or systematic confusion between distinct gestures, evidencing the effectiveness of the feature extraction approach and the dataset balancing applied during the preprocessing stage.

The confusion matrices complement the analysis by explicitly showing which gesture classes are confused with one another. In optimal scenarios (200/10), both KNN (Figure 5) and RF (Figure 6) exhibit practically perfect diagonals and residual confusions between gestures with similar muscular configurations, which is consistent with the leading role of these models throughout the experiments.

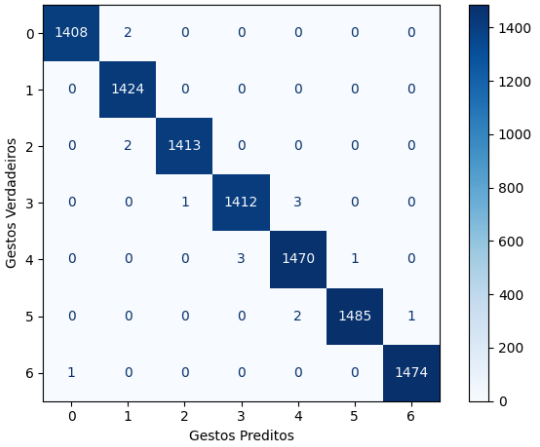

Figure 5. KNN confusion matrix with a 200-sample window and a stride of 10.

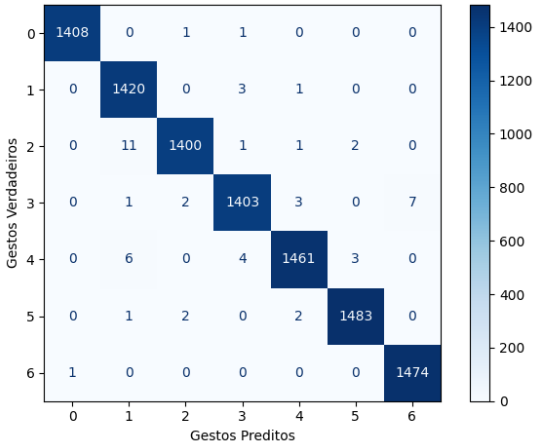

Figure 6. RF confusion matrix with a 200-sample window and a stride of 10.

When increasing the stride to 50 in the RF (Figure 7), a greater dispersion outside the diagonal is observed, indicating a loss of temporal resolution and a drop in performance, in line with the quantitative analysis for longer strides. This behavior reinforces the importance of smaller strides to maximize separability between classes in sEMG.

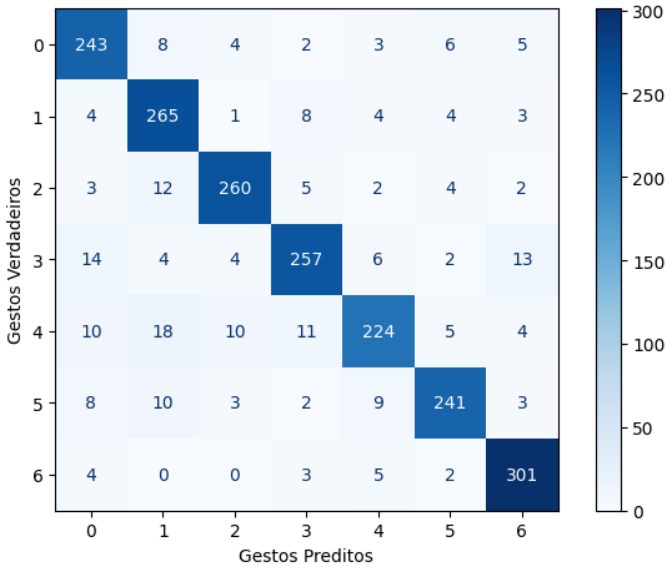

Figure 7. RF confusion matrix with a 200-sample window and a stride of 50.

4.2. Limitations

Despite the promising results, this study has some limitations. First, the overlapping sliding-window strategy may introduce optimism in the reported metrics, since adjacent windows share a large proportion of samples, and temporally close windows from the same gesture repetition may appear in both the training and test partitions. Although the stratified hold-out split preserves class proportions, future work should adopt a subject-

independent evaluation protocol — such as leave-one-subject-out cross-validation — to provide a more conservative and generalizable estimate of classifier performance.

Second, the numerical differences in accuracy observed between classifiers across configurations are, in some cases, small (e.g., KNN at 99.84% versus RF at 99.48% for the 200/10 configuration), and no formal statistical hypothesis testing was conducted to assess whether these differences are statistically significant. Applying non-parametric tests such as the Wilcoxon signed-rank test or the Friedman test across cross-validation folds would strengthen the conclusions drawn from the comparisons and is recommended for future extensions of this work.

Third, all experiments were conducted exclusively on data from healthy individuals. The generalization of the proposed pipeline to amputee users — where residual muscle activity and electrode placement differ substantially — remains an open and important challenge, which the authors intend to address in subsequent studies.

5. Conclusion

This work developed and evaluated, based on a set of experiments and a consolidated technical foundation, an approach for manual gesture recognition from surface electromyographic (sEMG) signals. Using the NinaPro database, we implemented and compared several machine learning algorithms, focusing on time-domain feature extraction and the impact of window size and temporal stride.

During the experimentation, seven distinct gestures were analyzed in a controlled environment, using signals from healthy individuals. The KNN, RF, SVM, and MLP classifiers yielded expressive results, with RF and SVM achieving accuracies above 99% in certain configurations. These results demonstrate the potential of supervised methods for pattern recognition in sEMG signals, despite the physiological and temporal variations inherent in such data.

It is worth noting, however, that the experiments were conducted exclusively on data from healthy individuals using a stratified hold-out split, which may introduce optimism in the reported metrics due to window overlap. Future work should validate the pipeline under subject-independent protocols and extend the evaluation to amputee populations.

With the consolidated methodology and the obtained results, this work is expected to contribute to the development of assistive systems based on myoelectric control, especially in scenarios where non-invasive, responsive, and low-cost solutions are desired. As a continuation of this study, we intend to apply the developed models to databases of amputee users, thereby expanding the algorithms' applicability to real-world use cases. Furthermore, deep learning techniques could be explored to improve classifier robustness and enable real-time detection of transitions between gestures. Another point to explore in future work is the analysis of feature redundancy and the possibility of reducing data dimensionality without significant performance loss. This could favor the deployment of lighter, more efficient models for embedded devices, such as robotic prostheses or portable rehabilitation systems.

Finally, it is hoped that this work will serve as a foundation for future research in bioengineering, assistive rehabilitation, and human-computer interaction, contributing to

both scientific advancement and the development of practical solutions that improve the quality of life for people with motor disabilities.

References

Atzori, M., Gijsberts, A., Heynen, S., Hager, A.-G. M., Deriaz, O., Van Der Smagt, P., Castellini, C., Caputo, B., and Müller, H. (2012). Building the ninapro database: A resource for the biorobotics community. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1258–1265. IEEE.

Corbett, E. A., Perreault, E. J., and Kuiken, T. A. (2011). Comparison of electromyography and force as interfaces for prosthetic control. *Journal of rehabilitation research and development*, 48(6):629–638.

De Luca, C. J. (2002). Surface electromyography: Detection and recording. Technical report, DelSys Incorporated, Boston, USA.

Garcia, M. C. and Vieira, T. (2011). Surface electromyography: Why, when and how to use it. *Revista andaluza de medicina del deporte*, 4(1):17–28.

Gopal, P., Gesta, A., and Mohebbi, A. (2022). A systematic study on electromyography-based hand gesture recognition for assistive robots using deep learning and machine learning models. *Sensors*, 22(10):3650.

Kadavath, M. R. K., Nasor, M., and Imran, A. (2024). Enhanced hand gesture recognition with surface electromyogram and machine learning. *Sensors*, 24(16):5231.

McDonald, C. L., Westcott-McCoy, S., Weaver, M. R., Haagsma, J., and Kartin, D. (2021). Global prevalence of traumatic non-fatal limb amputation. *Prosthetics and orthotics international*, 45(2):105–114.

Mills, K. R. (2005). The basics of electromyography. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii32–ii35.

Oskoei, M. A. and Hu, H. (2007). Myoelectric control systems—a survey. *Biomedical signal processing and control*, 2(4):275–294.

Phinyomark, A., Phukpattaranont, P., and Limsakul, C. (2012). Feature reduction and selection for emg signal classification. *Expert Systems with Applications*, 39(8):7420–7431.

Tanaka, T., Nambu, I., Maruyama, Y., and Wada, Y. (2022). Sliding-window normalization to improve the performance of machine-learning models for real-time motion prediction using electromyography. *Sensors*, 22(13):5005.