DOI: 10.37190/abb/221221

Multi-Dataset Benchmarking and Preliminary Transfer Learning for Lower-Limb EMG Classification Toward Adaptive Prosthesis Control

Jafar Jafar[1*], Cezary Rzymkowski[1,2], Andrzej Kordecki[1]

[1]Warsaw University of Technology, Warsaw, Poland

[2]Centre of Excellence for Health Technologies (HealthTEC), Poland

[*]Corresponding author: Jafar Jafar, Warsaw University of Technology, Warsaw, Poland, e-mail address: jafar.dokt@pw.edu.pl

Submitted: 19th March 2026

Accepted: 29th April 2026

Abstract

Purpose

Reliable decoding of lower-limb motor intent from surface electromyography signals remains challenging due to electrode placement variability, muscle fatigue, inter-subject differences, and differences between able-bodied and amputee populations. This study evaluates the robustness of commonly used machine-learning approaches for lower-limb electromyography classification under these sources of variability.

Methods

A unified benchmarking analysis was conducted across four lower-limb electromyography datasets: UCI Lower Limb Electromyography, Vastus Medialis Electromyography, K2MUSE, and the Above-Knee Sit-to-Stand amputee dataset. Three classification models—support vector machine, random forest, and multilayer perceptron—were evaluated using standardized preprocessing and feature extraction pipelines. Cross-subject robustness was assessed using leave-one-subject-out validation on the K2MUSE dataset. A preliminary transfer-learning experiment was also performed by fine-tuning a multilayer perceptron pretrained on able-bodied data using the amputee sit-to-stand dataset.

Results

The random forest model achieved the highest performance in datasets with lower variability, with macro F1 scores of 0.9033 ± 0.0048 on the UCI dataset and 0.6000 ± 0.0087 on the Vastus dataset. Under cross-subject non-ideal acquisition conditions, performance decreased for all models. The multilayer perceptron showed competitive robustness in the K2MUSE Ideal-versus-Fatigue LOSO evaluation (0.476 ± 0.058 macro-F1), while random forest showed the most compact subject-wise distribution (median 0.476 [IQR 0.049]). In the transfer-learning experiment, fine-tuning improved performance on the Above-Knee STS dataset relative to training from scratch (macro-F1 0.5540 ± 0.0141 vs. 0.5203 ± 0.0371).

Conclusions

These results highlight the different strengths of classical and neural machine-learning models under varying electromyography variability conditions and provide a foundation for more robust intention-decoding systems for adaptive lower-limb prosthetic control.

Keywords: Surface electromyography, lower-limb biomechanics, neuromuscular control, gait analysis, transfer learning, prosthesis control.

1. Introduction

Lower limb loss affects millions of individuals worldwide and is projected to increase substantially in the coming decades due to vascular disease, trauma, and complications of diabetes [1]. Above-knee (transfemoral) amputees face particularly demanding mobility challenges because the loss of both knee musculature and proprioceptive feedback profoundly impacts gait stability and limb control. Although modern microprocessor-controlled prosthetic knees provide improved safety and energy efficiency [2], these devices still rely on predefined mechanical modes rather than direct decoding of the user's neuromuscular intent. This gap motivates the development of intention-driven prosthesis controllers that can adapt to individual users and varying real-world conditions.

Surface electromyography (sEMG) provides a non-invasive window into muscle activation and motor commands and has been extensively explored as a control signal for prosthetic and assistive technologies [3-4]. sEMG-based intention decoding, however, remains hindered by several well-known sources of variability: electrode displacement, skin impedance changes, muscle fatigue, and inter-subject differences [5–8]. Recent systematic reviews on EMG-based lower-limb prosthetic control have similarly emphasized that signal variability, electrode placement, and methodological heterogeneity remain major obstacles to robust real-world deployment [9–10]. These sources of variability introduce *domain shifts*—changes in data distribution between training and deployment—that degrade model performance outside controlled laboratory environments.

To address these challenges, EMG research has evolved along two main trajectories. The first focuses on the design of robust feature extraction and classical machine-learning pipelines such as Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), and Random Forests (RF) [7,12]. These methods achieve strong performance under stable conditions but often struggle with cross-session or cross-subject generalization [13]. The second trajectory employs neural-network-based approaches such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-LSTM architectures [14–17], which learn hierarchical features and have shown improved adaptability. Transfer learning methods—including fine-tuning, domain-adversarial learning, and representation alignment—have demonstrated additional promise in reducing calibration requirements for new users or new recording conditions [15–17].

Despite these advances, several gaps remain in the lower-limb EMG literature. First, most studies focus on isolated datasets, limiting understanding of how models behave across different muscles, electrode configurations, movement tasks, or recording environments. Second, comparative evaluations between classical and neural classifiers are typically dataset-specific, making it difficult to derive general conclusions about robustness. Third, transfer learning between able-bodied and amputee populations—highly relevant for prosthesis control—remains underexplored, especially with lightweight models suitable for real-time embedded systems. This gap is particularly important in lower-limb prosthetics, where recent reviews have noted that robust EMG-driven control remains difficult to validate consistently across users, tasks, and hardware configurations [9–11].

In response to these gaps, this study presents a multi-dataset benchmarking analysis combined with a preliminary transfer-learning evaluation. Four publicly available lower-limb EMG datasets were selected to represent a spectrum of recording conditions and movement tasks: (1) UCI Lower Limb EMG [19], (2) Vastus Medialis EMG [20], (3) the K2MUSE multimodal lower-limb dataset [21], and (4) the Above-Knee Sit-to-Stand (STS) dataset featuring unilateral transfemoral amputees [22]. Together, these datasets encompass differences in electrode count (1–13 channels), task complexity (static posture, gait, incline walking), populations (able-bodied vs amputee), and acquisition noise (fatigue, electrode shift). Evaluating models across such heterogeneous datasets provides practical insights into the strengths and limitations of common EMG classification approaches.

Three machine-learning models were benchmarked: SVM, Random Forest (RF), and a tuned Multilayer Perceptron (MLP). The benchmarking examines within-dataset generalization using stratified holdout splits for UCI, Vastus, and K2MUSE; cross-session robustness using Leave-One-Subject-Out (LOSO) validation on K2MUSE; and a preliminary transfer-learning experiment in which an MLP pretrained on able-bodied K2MUSE data is fine-tuned on amputee STS EMG.

This approach yields two major contributions:

1) A unified evaluation of classical and neural classifiers across four heterogeneous EMG datasets, providing a clearer understanding of how model performance changes as dataset complexity increases.
2) A lightweight transfer-learning strategy demonstrating feasible adaptation from able-bodied to amputee EMG, offering an early step toward reducing calibration burden in prosthesis control.

The insights from this study contribute to the methodological foundation needed for more advanced multimodal work—such as EMG + mechanomyography (MMG) fusion and real-time prosthesis control—which will be the focus of future research.

2. Materials and Methods

2.1. Datasets

Four publicly available EMG datasets were selected to cover a range of muscles, motions, populations, and acquisition conditions: UCI Lower Limb EMG [19], Vastus Medialis EMG [20], the K2MUSE multimodal gait dataset [21], and the Above-Knee STS amputee dataset [22]. For clarity, the classification objective was explicitly defined for each dataset based on the original dataset annotations or condition labels. For the UCI Lower Limb and Vastus Medialis datasets, the task consisted of multi-class classification of basic postures or movement types using the original labels provided by the dataset authors. In contrast, the K2MUSE dataset was used to evaluate robustness to non-ideal acquisition conditions, where classification targets were defined according to recording conditions (Ideal, Fatigue, and Electrode Shift), resulting in both binary and three-class classification scenarios. Finally, the Above-Knee STS dataset was formulated as a binary classification task distinguishing stand-up and sit-down phases on the intact limb, consistent with the original dataset labelling. Table 1 lists the main characteristics and specifications of the EMG datasets analyzed in this study. These datasets vary in channel counts (1–13 electrodes), sampling rates (1–2 kHz), subject populations (able-bodied vs amputee), and task complexity (static postures, gait phases, incline walking, sit-to-stand transitions). A standardized preprocessing pipeline was applied across all datasets to ensure comparability.

Table 1. Dataset Characteristics and Specifications.

Property	UCI Lower Limb Dataset [19]	Vastus Medialis Dataset [20]	K2MUSE dataset [21]	Above-Knee STS Dataset [22]
Subjects	22 males (11 normal and 11 (knee abnormalities))	22 healthy (11 males and 11 females)	30 healthy (20 males, 10 females)	9 unilateral above-knee (transfemoral) amputees (K3-level community ambulators)
EMG Electrodes	4 (RF, BF, VM, ST)	1 (VM)	13 (TA, MG, LG, SOL, RF, VL, VM, BF, SEM)	4 surface EMG sensors (Delsys Trigno Avanti) placed on the intact-side lower limb (non-prosthesis side)
Sampling Frequency	1 kHz	1 kHz	1 kHz	2 kHz
Recording Duration	5 repetitions	5 seconds, 3 repetitions	40 seconds, 5 repetitions	8-10 repetitions
Classification target (number of classes)	Posture/activity classification: sitting, standing,	Movement type classification: sitting, standing,	Acquisition-condition classification	Binary task-phase classification: intact-side stand-

	walking (3 classes, original labels)	gait (3 classes, original labels)	under locomotion: Ideal vs Fatigue (2 classes); Ideal vs Shift (2 classes); Ideal/Fatigue/Shift (3 classes, derived from dataset condition labels)	up vs sit-down (2 classes, original labels)
Original Preprocessing	Band pass filter	Band pass filter	Band pass filter, notch filter, rectified and segmented, time normalization, stored values per channel	segmentation of start/stop indices using knee angle & velocity thresholds.

Note: Muscle abbreviations: RF (Rectus Femoris), BF (Biceps Femoris), VM (Vastus Medialis), ST (Semitendinosus), TA (Tibialis Anterior), MG (Medial Gastrocnemius), LG (Lateral Gastrocnemius), SOL (Soleus), VL (Vastus Lateralis), SEM (Semimembranosus).

2.2.Preprocessing

As shown in Figure 1, raw EMG signals are first band-pass filtered to remove motion artifacts and high-frequency noise, followed by rectification to preserve signal magnitude. The rectified EMG is then segmented using a sliding window of 200 ms with 50% overlap to balance temporal resolution and signal stationarity. Within each window, standard time-domain features—root mean square (RMS), mean absolute value (MAV), zero crossing (ZC), and waveform length (WL)—are extracted on a per-channel basis. For datasets involving multiple recording channels, channel-wise features are subsequently aggregated using mean and standard deviation pooling, resulting in a compact and interpretable feature vector of eight features per window suitable for machine-learning–based classification.

A common preprocessing pipeline was used to ensure that performance differences were primarily attributable to model behaviour and dataset characteristics rather than to dataset-specific signal processing choices. We acknowledge that dataset-specific preprocessing may improve absolute performance for individual datasets, particularly given differences in hardware, channel configuration, and recording noise. However, such customization would reduce methodological consistency and make cross-dataset comparison less interpretable. Thus, the standardized pipeline was chosen as a benchmarking strategy focused on comparability and reproducibility.

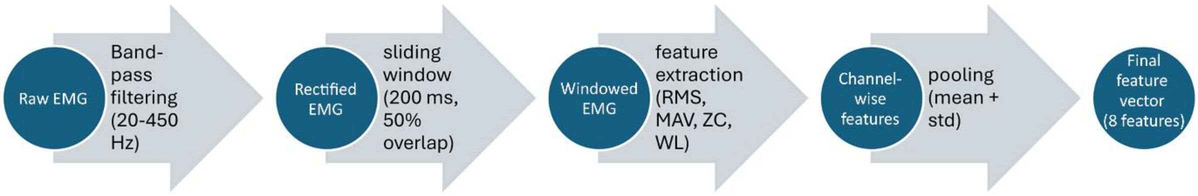

Figure 1. Standardized EMG preprocessing and feature extraction pipeline used across all datasets.

2.3.Feature Extraction

Time-domain features, including root mean square (RMS), mean absolute value (MAV), zero crossing (ZC), and waveform length (WL), were extracted from each EMG channel using a sliding window of 200 ms with 50% overlap as shown in Figure 1. Feature extraction was performed on rectified EMG signals to characterize muscle activation intensity and temporal complexity within each window. For multi-channel datasets (K2MUSE and Above-Knee), channel-wise features were subsequently aggregated using mean and standard deviation pooling across channels, capturing both global activation levels and inter-channel variability. This pooling strategy yielded a compact feature representation consisting of eight features per window, which was used as input for all classification models.

The selected feature set was intended as a compact and standardized time-domain baseline rather than an optimized feature subset for each dataset. RMS, MAV, ZC, and WL were chosen because they are widely used in EMG pattern recognition, computationally efficient, and physiologically interpretable, capturing signal amplitude, activation complexity, and temporal variation within each analysis window. This choice was made to preserve methodological consistency across heterogeneous datasets and to support future translation to lightweight real-time prosthetic control systems. We acknowledge, however, that other feature families, including frequency-domain, time-frequency, or learned representations, may provide additional discriminative information. A systematic comparison of feature sets and feature-importance analysis was beyond the scope of the present benchmark and should be addressed in future work.

2.4.Machine Learning Models

Three supervised machine-learning classifiers were employed to evaluate lower-limb EMG classification performance under different levels of variability. A Support Vector Machine (SVM) with a radial basis function (RBF) kernel was used as a classical margin-based baseline, with class weighting applied to mitigate class imbalance. A Random Forest (RF) classifier consisting of 200 decision trees with a maximum depth of 10 was included to represent ensemble-based methods that are commonly used in EMG pattern recognition due to their robustness and interpretability. In addition, a lightweight Multilayer Perceptron (MLP) was implemented to assess the potential benefits of nonlinear neural representations while maintaining computational efficiency. The MLP architecture comprised three hidden layers with 128, 64, and 32 units, respectively, using ReLU activation functions. Model training was performed using the Adam optimizer with early stopping to prevent overfitting, and class weighting was applied to address class imbalance. While SVM and RF were evaluated only in within-dataset and cross-subject benchmarking scenarios, the MLP was additionally used in the transfer-learning experiments due to its suitability for weight initialization and fine-tuning across datasets. All models were trained and evaluated using the dataset-specific validation protocols described in Section 2.5.

2.5.Evaluation Protocol

Evaluation protocols were selected in a dataset-specific manner to reflect differences in subject availability, recording conditions, and the intended robustness analyses as illustrated in Figure 2. For the UCI Lower Limb and Vastus Medialis datasets, baseline model performance was evaluated using stratified 80/20 train–test splits. This protocol was chosen due to the limited number of subjects and the absence of explicit session or condition labels, making subject-wise cross-validation impractical.

For the K2MUSE dataset, leave-one-subject-out (LOSO) validation was employed to explicitly assess cross-subject robustness under non-ideal acquisition conditions, including muscular fatigue and electrode shift. In this setting, models were trained on data from all but one subject and evaluated on the held-out subject, enabling direct quantification of inter-subject generalization. LOSO analysis was applied exclusively to the K2MUSE dataset because it is the only dataset in this study that provides sufficient subject count and structured condition labels to support meaningful subject-wise validation.

For the Above-Knee sit-to-stand dataset, a transfer-learning protocol was adopted. An MLP model was first pretrained on the K2MUSE dataset and subsequently fine-tuned on the Above-Knee data using an 80/20 split. Model performance was evaluated on the held-out portion of the target dataset using accuracy and macro-F1 metrics. LOSO validation was not applied to this dataset due to the limited number of amputee subjects ($n = 9$), which would lead to unstable performance estimates. For comparison purposes, an MLP trained from scratch on the Above-Knee dataset using the same architecture was included as a target-domain baseline. Both the scratch and fine-tuned models were evaluated under repeated stratified hold-out runs to enable a direct comparison of transfer-learning effects.

Although a unified validation protocol across all datasets would be preferable for strict numerical comparison, it was not feasible because the datasets differ substantially in subject availability, annotation structure, recording conditions, and intended analytical use. Therefore, each protocol was selected to answer a different methodological question: baseline within-dataset classification for the UCI Lower Limb and Vastus Medialis datasets, cross-subject robustness under non-ideal acquisition conditions for K2MUSE, and cross-population adaptation with limited target data for the Above-Knee STS dataset. For this reason, the results should not be interpreted as direct one-to-one comparisons of absolute performance across all datasets. Instead, they are intended to illustrate how model behaviour changes as variability and domain shift increase across progressively more challenging scenarios.

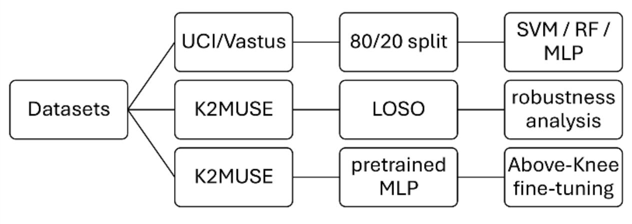

Figure 2. Evaluation protocols used for baseline classification, LOSO robustness analysis, and transfer learning across datasets.

3. Results

3.1. Baseline Performance on UCI EMG

As shown in Figure 3, repeated stratified hold-out evaluation across 20 runs confirmed that Random Forest (RF) consistently achieved the highest performance on the UCI dataset, with an accuracy of 0.9133 ± 0.0043 and a macro-F1 score of 0.9033 ± 0.0048. In comparison, the Multilayer Perceptron (MLP) reached an accuracy of 0.8721 ± 0.0123 and a macro-F1 of 0.8569 ± 0.0133, whereas the Support Vector Machine (SVM) showed substantially lower performance, with an accuracy of 0.7346 ± 0.0087 and a macro-F1 of 0.7065 ± 0.0098. These results indicate that RF not only provides the best mean performance (0.9133 ± 0.0043 accuracy; 0.9033 ± 0.0048 macro-F1) but also exhibits lower variability than MLP (0.8721 ± 0.0123; 0.8569 ± 0.0133) and SVM (0.7346 ± 0.0087; 0.7065 ± 0.0098).

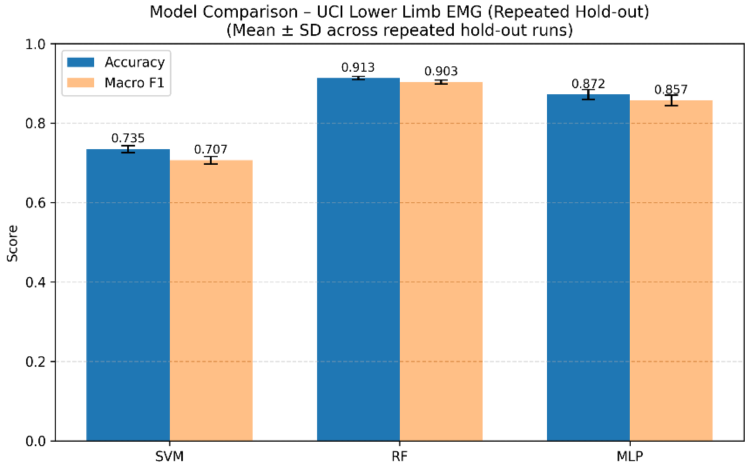

Figure 3. UCI EMG classification performance under 20 repeated stratified hold-out runs. Bars show mean accuracy and macro-F1; error bars indicate standard deviation.

3.2.Baseline Performance on Vastus EMG

As illustrated in Figure 4, repeated stratified hold-out evaluation across 20 runs showed that classification performance on the Vastus dataset was consistently lower than that observed for the UCI dataset. Random Forest (RF) again achieved the best overall performance, with an accuracy of 0.6094 ± 0.0093 and a macro-F1 score of 0.6000 ± 0.0087. The Multilayer Perceptron (MLP) yielded lower mean performance, with an accuracy of 0.5232 ± 0.0155 and a macro-F1 of 0.5103 ± 0.0147, while Support Vector Machine (SVM) produced the weakest results, with an accuracy of 0.5022 ± 0.0096 and a macro-F1 of 0.4943 ± 0.0098. Nevertheless, RF remained the most reliable classifier across repeated splits, with 0.6094 ± 0.0093 accuracy and 0.6000 ± 0.0087 macro-F1, compared with 0.5232 ± 0.0155 and 0.5103 ± 0.0147 for MLP, and 0.5022 ± 0.0096 and 0.4943 ± 0.0098 for SVM.

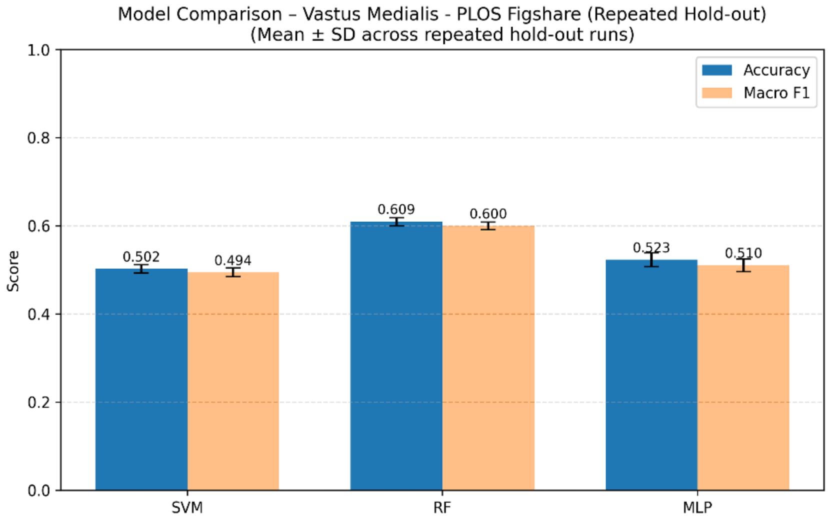

Figure 4. Vastus EMG classification performance under 20 repeated stratified hold-out runs. Bars show mean accuracy and macro-F1; error bars indicate standard deviation.

3.3. K2MUSE LOSO - Acquisition Condition Classification (Ideal vs Fatigue)

As shown in Figure 5, subject-wise LOSO evaluation on the K2MUSE dataset revealed substantial inter-subject variability in discriminating Ideal versus Fatigue conditions. The MLP achieved the highest mean macro-F1 (0.476 ± 0.058), followed closely by RF (0.470 ± 0.033), whereas SVM produced the lowest mean macro-F1 (0.430 ± 0.068). However, the median-based summary showed that RF had the highest central tendency, with a median macro-F1 of 0.476 and an interquartile range (IQR) of 0.049, compared with 0.464 [0.083] for MLP and 0.419 [0.082] for SVM. These results indicate that RF yielded the most compact and stable subject-wise performance distribution, with a median macro-F1 of 0.476 and an IQR of 0.049, whereas MLP showed a comparable mean macro-F1 (0.476 ± 0.058) but a broader median spread (0.464 [0.083]). In contrast, SVM remained lower overall, with a mean macro-F1 of 0.430 ± 0.068 and a median of 0.419 [0.082]. Paired Wilcoxon signed-rank testing further showed that both RF and MLP significantly outperformed SVM in macro-F1 (RF vs SVM: p = 0.0277; MLP vs SVM: $p < 0.001$), whereas the difference between RF and MLP was not statistically significant (p = 0.9515). Overall, these findings suggest that fatigue introduces meaningful cross-subject variability, but both RF and MLP provide stronger subject-generalization performance than SVM under this condition.

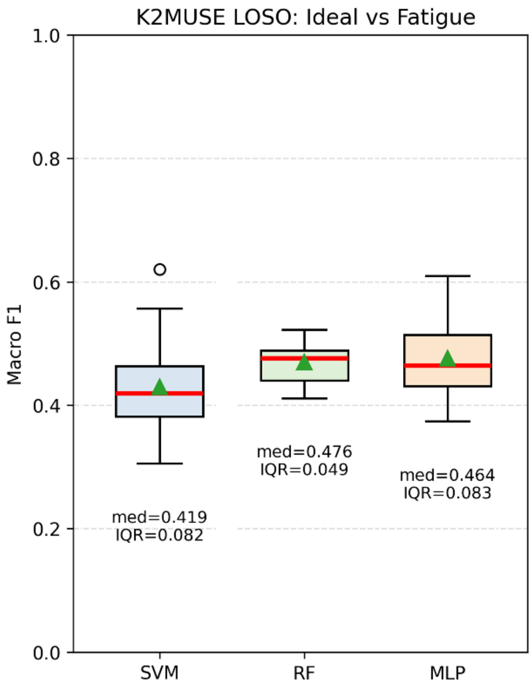

Figure 5. LOSO subject-wise macro-F1 distributions for K2MUSE Ideal versus Fatigue classification. Annotations indicate median and IQR.

3.4. K2MUSE LOSO - Acquisition Condition Classification (Ideal vs Electrode Shift)

As shown in Figure 6, LOSO evaluation for the Ideal versus Shift condition revealed pronounced subject-dependent variability across all classifiers, confirming electrode shift as a major source of instability in the K2MUSE dataset. Random Forest (RF) achieved the highest mean macro-F1 (0.633 ± 0.246) and the highest mean accuracy (0.913 ± 0.129), whereas SVM and MLP showed similar but lower average macro-F1 values of 0.474 ± 0.056 and 0.473 ± 0.114, respectively. However, the subject-wise distribution for RF was highly dispersed, with a median macro-F1 of only 0.499 and a very large interquartile range (IQR = 0.531), compared with 0.454 [0.041] for SVM and 0.457 [0.086] for MLP. This broad spread is partly explained by the fact that several held-out subjects yielded macro-F1 = 1.0 for RF, causing the upper part of the boxplot to reach the ceiling of the metric range. Accordingly, RF obtained the highest mean macro-F1 (0.633 ± 0.246), but this performance was not consistently distributed across all subjects, as reflected by its median of only 0.499 and a very large IQR of 0.531. By comparison, SVM and MLP showed lower mean macro-F1 values of 0.474 ± 0.056 and 0.473 ± 0.114, with median [IQR] values of 0.454 [0.041] and 0.457 [0.086], respectively. Paired Wilcoxon signed-rank testing showed that RF significantly outperformed SVM (p = 0.0310) and MLP (p = 0.0164) in macro-F1, while no significant difference was observed between SVM and MLP (p = 0.9354). Overall, these findings indicate that electrode shift induces strong domain variability, and although RF can perform extremely well for some subjects, including several held-out cases with macro-F1 = 1.0, cross-subject robustness under this condition remains limited and uneven across the dataset.

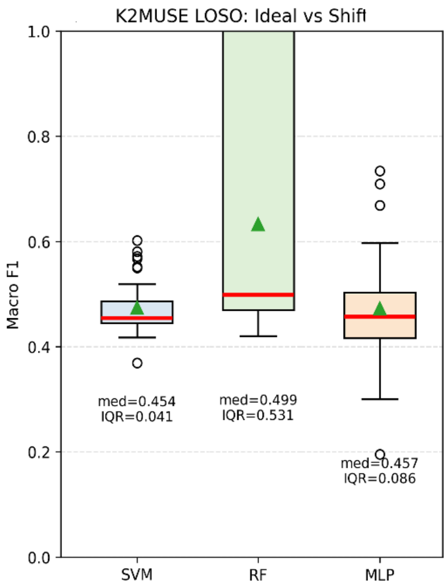

Figure 6. LOSO subject-wise macro-F1 distributions for K2MUSE Ideal versus Shift classification. Annotations indicate median and IQR.

3.5. K2MUSE LOSO - Three-Class Acquisition Condition Classification (Ideal/Fatigue/Shift)

As shown in Figure 7, the three-class LOSO classification task on the K2MUSE dataset was more challenging than the corresponding binary condition analyses, resulting in lower macro-F1 values across all models. Among the evaluated classifiers, the MLP achieved the highest overall performance, with a mean macro-F1 of 0.349 ± 0.073 and a median of 0.353 [0.056]. Random Forest (RF) ranked second, yielding a mean macro-F1 of 0.312 ± 0.033 and a median of 0.306 [0.028], while Support Vector Machine (SVM) showed the lowest performance, with 0.282 ± 0.047 and 0.277 [0.073], respectively. These results suggest that the MLP provides better discrimination than the conventional classifiers, with a mean macro-F1 of 0.349 ± 0.073 and a median of 0.353 [0.056], compared with 0.312 ± 0.033 and 0.306 [0.028] for RF, and 0.282 ± 0.047 and 0.277 [0.073] for SVM. Overall, the three-class setting highlights the increased difficulty of generalized acquisition-condition recognition and supports the use of MLP as the most effective model among those evaluated in this scenario.

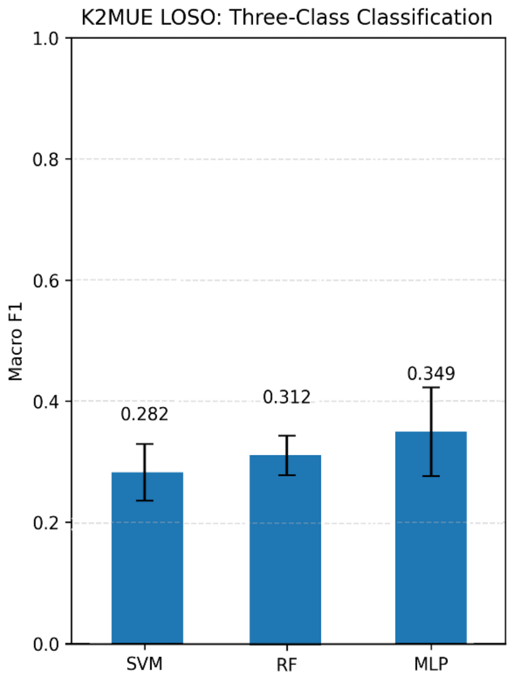

Figure 7. LOSO performance for K2MUSE three-class acquisition-condition classification (Ideal, Fatigue, Shift). Bars show mean macro-F1; error bars indicate standard deviation.

3.6. Transfer Learning: K2MUSE to Above-Knee STS

To directly assess the benefit of transfer learning, the Multilayer Perceptron (MLP) used in this experiment was initialized with weights pretrained on the K2MUSE dataset using the Ideal-versus-Fatigue acquisition-condition task and then fine-tuned on the Above-Knee STS dataset.

Its performance was compared against an identical MLP trained from scratch on the target dataset using the same architecture. To provide a more reliable comparison, both settings were evaluated under 10 repeated stratified hold-out runs.

As shown in Figure 8, the fine-tuned MLP consistently outperformed the scratch baseline on the Above-Knee STS dataset. Fine-tuning achieved an accuracy of 0.5572 ± 0.0126 and a macro-F1 score of 0.5540 ± 0.0141, whereas training from scratch reached 0.5396 ± 0.0137 and 0.5203 ± 0.0371, respectively. The improvement was statistically significant according to paired Wilcoxon signed-rank testing for both accuracy and macro-F1 ($p = 0.0039$). In addition to improving the mean performance, fine-tuning also reduced the variability of macro-F1 across repeated splits, indicating a more stable initialization for the target task. Although the absolute gain remained modest, these results suggest that representations learned from a related EMG acquisition-condition task can be beneficially transferred to the Above-Knee STS classification problem under limited target-data conditions. Because the Above-Knee STS dataset includes only nine transfemoral amputee subjects, these transfer-learning results should be interpreted cautiously and viewed as preliminary rather than fully generalizable.

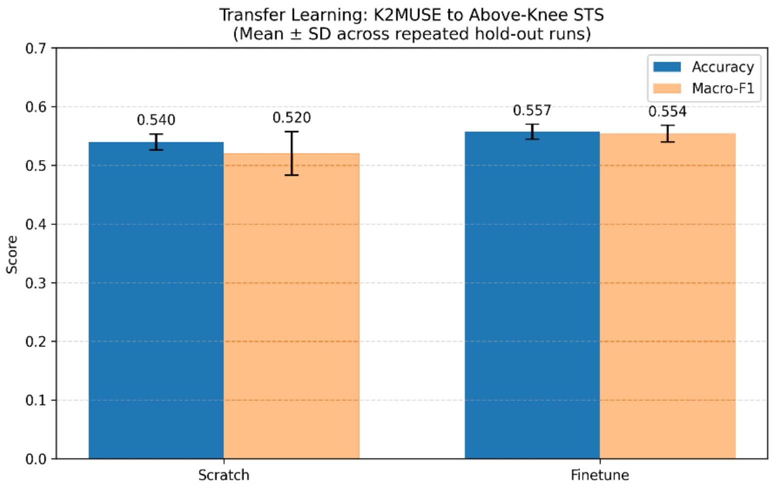

Figure 8. Transfer-learning performance on the Above-Knee STS dataset under 10 repeated stratified hold-out runs. Bars show mean accuracy and macro-F1; error bars indicate standard deviation.

3.7.Error Analysis

Error analysis indicated that the most difficult predictions occurred under non-ideal and cross-domain conditions rather than in the baseline datasets. In K2MUSE, subject-wise performance variability under fatigue and electrode shift suggests that these conditions increased overlap between class distributions and reduced robustness across held-out subjects. The three-class acquisition-condition task was particularly challenging, reflecting greater confusion when Ideal, Fatigue, and Shift conditions had to be separated simultaneously. In the transfer-learning setting, the confusion matrix of the fine-tuned MLP showed that both intact and prosthetic trials were still misclassified in a non-negligible number, despite the overall improvement over training from scratch. This indicates that transfer learning improved separability but did not eliminate the domain mismatch between able-bodied and amputee EMG patterns.

4. Discussion

This study establishes a multi-dataset benchmark for lower-limb EMG classification and demonstrates the feasibility of cross-population transfer learning. Classical models such as RF excel in controlled conditions, while MLP demonstrates improved robustness under specific sources of variability, particularly electrode shift. Transfer learning offers a promising pathway for personalized prosthesis control with limited amputee training data.

The baseline results on the UCI and Vastus Medialis datasets demonstrate that classical ensemble-based methods, particularly Random Forest, perform strongly under relatively stable recording conditions. In these datasets, EMG signals are characterized by limited inter-session variability, constrained task definitions, and consistent electrode placement. Under such conditions, tree-based models effectively exploit discriminative time-domain features, yielding high accuracy and macro-F1 scores. The consistently weaker performance of SVM suggests that margin-based classifiers are more sensitive to overlapping feature distributions, particularly when class separability is limited by single-muscle recordings, as observed in the Vastus dataset. Under such conditions, tree-based models effectively exploit discriminative time-domain features, yielding high accuracy and macro-F1 scores. This pattern is consistent with previous EMG pattern-recognition studies showing that conventional feature-based classifiers perform strongly in controlled settings but often lose robustness when variability increases across sessions or subjects [7][12–13].

In contrast, the K2MUSE LOSO experiments highlight the limitations of classical models when exposed to substantial physiological and acquisition-induced variability. Across both Ideal–Fatigue and Ideal–Shift conditions, performance degradation is evident for all classifiers, confirming muscular fatigue and electrode displacement as dominant sources of domain shift in lower-limb EMG. Notably, the MLP demonstrates more stable performance across subjects, with narrower interquartile ranges in fatigue conditions and higher median macro-F1 under electrode shift. This behaviour suggests that even relatively shallow neural networks can learn representations that are more resilient to non-stationary EMG characteristics than fixed decision boundaries, in agreement with earlier studies reporting improved adaptability of neural and transfer-learning approaches under domain shift [13][17].

One possible explanation is that electrode shift perturbs EMG feature distributions in a nonlinear and subject-dependent way. The MLP may better accommodate such perturbations because it learns flexible nonlinear feature transformations, whereas Random Forest depends on fixed threshold-based partitions that can become unstable when the input distribution shifts. This may explain why RF achieved a high mean score but showed highly uneven performance across subjects under electrode shift. Recent work has likewise shown that Random Forest-based myoelectric controllers can be made more robust to posture-related distribution shifts when adaptive or self-calibrating mechanisms are introduced, suggesting that the instability observed in the present benchmark may reflect limitations of static RF decision boundaries rather than of tree-based methods in general [23–24].

The three-class LOSO results further emphasize the challenge of realistic EMG-based state classification. Simultaneously discriminating between Ideal, Fatigue, and Shift conditions leads to substantial performance drops across all models, with macro-F1 values remaining below 0.4. This outcome reflects pronounced overlap in EMG feature space when multiple physiological and sensor-related perturbations coexist. These findings underscore that robustness observed in binary classification does not directly extend to multi-state scenarios, reinforcing the need for adaptive or domain-aware modelling strategies in practical prosthetic control systems.

The transfer-learning experiment from K2MUSE to the Above-Knee STS dataset provides initial evidence that EMG representations learned from able-bodied populations can be partially transferred to amputee data. In repeated stratified hold-out evaluation, fine-tuning outperformed training from scratch, improving accuracy from 0.5396 ± 0.0137 to 0.5572 ± 0.0126 and macro-F1 from 0.5203 ± 0.0371 to 0.5540 ± 0.0141. This improvement was statistically significant for both metrics ($p = 0.0039$) and indicates that lightweight transfer learning can provide a more stable initialization for the target task under limited-data conditions. However, the confusion matrix reveals persistent misclassifications, reflecting residual domain mismatch arising from differences in neuromuscular organization, biomechanics, and recording context between able-bodied and transfemoral amputee subjects. An additional limitation of this study is the small size of the amputee dataset, which included only nine transfemoral subjects. As a result, the transfer-learning findings should be interpreted as preliminary and require validation in larger and more diverse amputee cohorts. This interpretation is consistent with earlier studies reporting improved adaptability of neural and transfer-learning approaches under domain shift [14,17], as well as with recent reviews highlighting that EMG-driven control in lower-limb prostheses remains constrained by signal instability, user variability, and the limited availability of standardized validation frameworks, particularly in amputee populations [9–11].

Collectively, these results suggest a clear hierarchy of model suitability: classical methods remain effective in controlled, low-variability settings, whereas neural models offer improved robustness as domain shift increases. Importantly, even simple neural architectures can provide meaningful gains without the computational overhead of deep end-to-end models, making them attractive for embedded prosthetic applications.

Building on the insights of this study, several directions for future research are identified. First, integrating mechanomyography (MMG) with EMG signals may improve robustness to electrode-related artifacts and physiological variability, directly aligning with the author’s ongoing PhD research. Second, more advanced deep-learning architectures, including CNN–LSTM and Transformer-based models, may better capture temporal dependencies and nonlinear signal variations than the lightweight MLP used in the present study. Such models have shown increasing promise in recent EMG and lower-limb movement-analysis research. Third, advanced domain adaptation techniques, including adversarial learning and feature alignment, and transfer-learning strategies designed for distribution shift, should be explored to further reduce cross-subject and cross-population mismatch. Finally, implementing and evaluating these models in real-time embedded systems will be critical for translating laboratory findings into clinically viable adaptive prosthetic controllers.

5. Conclusions

This study establishes a comprehensive multi-dataset benchmark for lower-limb EMG classification, systematically evaluating classical and neural machine-learning models under increasing levels of variability and domain shift. Results demonstrate that Random Forest achieves superior performance in stable, low-variability datasets, while Multilayer Perceptron models exhibit greater robustness in cross-subject and non-ideal acquisition conditions. Leave-One-Subject-Out analyses confirm that fatigue and electrode shift substantially degrade performance, particularly in multi-class scenarios, highlighting key challenges for real-world deployment.

Furthermore, the preliminary transfer-learning results indicate that fine-tuning models pretrained on able-bodied EMG data can improve classification performance in amputee datasets, supporting the feasibility of cross-population adaptation. Together, these findings provide a principled foundation for developing robust, adaptive EMG-based intent decoding systems and inform model selection strategies for future prosthetic control applications.

6. Acknowledgments

The authors would like to thank the maintainers of the publicly available datasets used in this study, including the UCI Lower Limb EMG dataset, the Vastus Medialis EMG dataset, the

K2MUSE dataset, and the Above-Knee Sit-to-Stand dataset, whose open access resources enabled the multi-dataset benchmarking conducted in this work.

The authors also gratefully acknowledge the support and guidance provided by the research supervisor and colleagues within the laboratory, whose discussions and feedback contributed to improving the methodological design and interpretation of the results.

This work was conducted as part of the author's doctoral research on neuromuscular sensing and adaptive prosthetic control. The authors further acknowledge the broader research community in biomedical engineering and prosthetics for their continued efforts in advancing open datasets, algorithms, and tools that support reproducible research.

References

[1] Ziegler-Graham K., MacKenzie E.J., Ephraim P.L., Travison T.G., Brookmeyer R., *Estimating the prevalence of limb loss in the United States: 2005 to 2050*, Arch Phys Med Rehabil, 2008, 89(3):422-429.

[2] Highsmith M.J., Kahle J.T., Bongiorni D.R., Sutton B.S., Groer S., Kaufman K.R., *Safety, energy efficiency, and cost efficacy of the C-Leg for transfemoral amputees: a review of the literature*, Prosthet Orthot Int, 2010, 34(4):362-377.

[3] De Luca C.J., *Surface electromyography: detection and recording*, Delsys Inc., 2002.

[4] Farina D., Jiang N., Rehbaum H., Holobar A., Graimann B., Dietl H., Aszmann O.C., *The extraction of neural information from the surface EMG for the control of upper-limb prostheses: emerging avenues and challenges*, IEEE Trans Neural Syst Rehabil Eng, 2014, 22(4):797-809.

[5] Clancy E.A., Morin E.L., Merletti R., *Sampling, noise-reduction and amplitude estimation issues in surface electromyography*, J Electromyogr Kinesiol, 2002, 12(1):1-16.

[6] Huang H., Zhang F., Hargrove L.J., Dou Z., Rogers D.R., Englehart K.B., *Continuous locomotion-mode identification for prosthetic legs based on neuromuscular–mechanical fusion*, IEEE Trans Biomed Eng, 2011, 58(10):2867-2875.

[7] Scheme E., Englehart K., *Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use*, J Rehabil Res Dev, 2011, 48(6):643-660.

[8] Young A.J., Hargrove L.J., Kuiken T.A., *The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift*, IEEE Trans Biomed Eng, 2011, 58(9):2537-2544.

[9] Cimolato A., Driessen J.J.M., Mattos L.S., De Momi E., Laffranchi M., De Michieli L., *EMG-driven control in lower limb prostheses: a topic-based systematic review*, J Neuroeng Rehabil, 2022, 19(1):43. https://doi.org/10.1186/s12984-022-01019-1.

[10] Ahkami B., Ahmed K., Thesleff A., Hargrove L., Ortiz-Catalan M., *Electromyography-based control of lower limb prostheses: a systematic review*, IEEE Trans Med Robot Bionics, 2023, 5(3):547-562. https://doi.org/10.1109/TMRB.2023.3282325.

[11] Le D., Cheng S., Gregg R.D., Ghaffari M., *Transfer learning for efficient intent prediction in lower-limb prosthetics: a strategy for limited datasets*, IEEE Robot Autom Lett, 2024, 9(5):4321-4328. https://doi.org/10.1109/LRA.2024.3379800.

[12] He J., Zhang D., Jiang N., Sheng X., Farina D., Zhu X., *User adaptation in long-term, open-loop myoelectric training: implications for EMG pattern recognition in prosthesis control*, J Neural Eng, 2015, 12(4):046005.

[13] Atzori M., Cognolato M., Müller H., *Deep learning with convolutional neural networks applied to electromyography data: a resource for the classification of movements for prosthetic hands*, Front Neurorobot, 2016, 10:9.

[14] Côté-Allard U., Fall C.L., Drouin A., Campeau-Lecours A., Gosselin C., Laviolette F., Gosselin B., *Deep learning for electromyographic hand gesture signal classification using transfer learning*, IEEE Trans Neural Syst Rehabil Eng, 2019, 27(4):760-771.

[15] Weiss K., Khoshgoftaar T.M., Wang D., *A survey of transfer learning*, J Big Data, 2016, 3(1):1-40.

[16] Du Y., Jin W., Wei W., Hu Y., Geng W., *Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation*, Sensors, 2017, 17(3):458.

[17] Ameri A., Akhaee M.A., Scheme E., Englehart K., *Regression convolutional neural network for improved simultaneous EMG control*, J Neural Eng, 2019, 16(3):036015.

[18] Chen X., Zhang D., Zhu X., *Application of a self-enhancing classification method to electromyography pattern recognition for multifunctional prosthesis control*, Front Neurol, 2022, 13:853207.

[19] Sanchez O., Sotelo J., *EMG dataset in lower limb*, UCI Machine Learning Repository, 2014. https://archive.ics.uci.edu/dataset/278/emg+dataset+in+lower+limb. Accessed 2025.

[20] Zhang Y., *Original EMG datasets*, PLoS One, 2017. https://figshare.com/articles/dataset/Extracting_time-frequency_feature_of_single-channel_vastus_medialis_EMG_signals_for_knee_exercise_pattern_recognition/5191081?file=8865790. Accessed 2025.

[21] Li J., Zhang B., Tan X., Chen W., Liu Z., Zhang J., Huo W., Huang J., Liu L., Zhao X., *K2MUSE: A human lower limb multimodal dataset*, Kaggle Dataset, 2025. https://www.kaggle.com/datasets/98d67c253a7c820668aed0690cae20343481b8f8f8e0dafbe93b0c76d91f0ce6. Accessed 2025.
[22] Hunt G., Foreman K.B., Lenzi T., Gabert L., Hansen C., *Transfemoral amputees stand-up and sit-down with their prescribed microprocessor prostheses*, Figshare Dataset, 2024. https://figshare.com/articles/dataset/Transfemoral_amputees_stand-up_and_sit-down_with_their_prescribed_microprocessor_prosthesis_Full-body_motion_capture_and_intact-side_lower_limb_EMGs/26282872. Accessed 2025.
[23] Jiang X., Ma C., Nazarpour K., *Posture-invariant myoelectric control with self-calibrating random forests*, Front Neurorobot, 2024, 18:1462023. https://doi.org/10.3389/fnbot.2024.1462023.
[24] Tanaka T., Nambu I., Wada Y., *Mitigating the impact of electrode shift on classification performance in electromyography applications using sliding-window normalization*, Sensors, 2025, 25:4119. https://doi.org/10.3390/s25134119.