WASSL-EGR : Wavelet Attention and State-Space Learning for Robust EMG-based Gesture Recognition

Mainak Ghosh
Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, India
mainakcgeccse@gmail.com

Anup Nandy
Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, India
nandy.anup@gmail.com

***Abstract*—Surface electromyography (sEMG) signals enable intuitive human–machine interaction by capturing muscle activation patterns associated with hand gestures. However, the accurate recognition of hand gestures using sEMG signals still remains challenging due to the complex nature of the signal variations. This work proposes a wavelet-based deep learning framework for EMG gesture recognition using time–frequency representations. The segmented sEMG signal is converted into the Continuous Wavelet Transform (CWT) spectrogram to obtain the multi-scale muscle activation. Two lightweight deep learning frameworks are developed to effectively learn gesture representations from these wavelet features. The Wavelet Attention Convolutional Neural Network (WA-CNN) integrates scale and channel attention to emphasize informative frequency bands and EMG electrodes. The Wavelet State-Space Model (WSSM) incorporates efficient temporal modeling of gesture dynamics. Experimental results on the Ninapro DB2 dataset demonstrate that the proposed models outperform conventional machine learning and deep learning baselines. In particular, WSSM achieves 95% accuracy while maintaining low model complexity suitable for real-time wearable EMG-based interaction systems.**

***Index Terms*—Electromyography (EMG), Gesture Recognition, Continuous Wavelet Transform, Attention Mechanism, State-Space Models, Human–Machine Interaction**

I. Introduction

Surface electromyography (sEMG) signals are an important tool in the direct interface and interpretation of muscle activities, facilitating human-machine interaction [1], [2]. sEMG is popularly used in prosthetics, rehabilitation, and assistive robots. However, EMG signals are highly non-stationary and exhibit strong variability across subjects, electrodes, and muscle activation patterns [3]. Due to this, reliable EMG-based gesture recognition remains a challenging task.

Conventional approaches in EMG signal recognition rely on handcrafted features extracted from the time or frequency domain [4]. Although these features are computationally efficient and summarize signal statistics, they often fail to capture the complex temporal dynamics of muscle activation [5]. Recent deep learning approaches attempt to learn representations directly from EMG signals [6]. However, models trained on raw signals may not effectively use the underlying time–frequency characteristics where gesture patterns are embedded.

To address these challenges, this work proposes a wavelet-based deep learning framework for EMG gesture recognition using the Ninapro DB2 dataset. EMG signals are segmented and transformed into time–frequency representations using Continuous Wavelet Transform (CWT), which enables multi-resolution analysis of transient muscle activation patterns. Two lightweight architectures are developed to learn gesture representations from the wavelet spectrograms. The Wavelet Attention Convolutional Neural Network (WA-CNN) integrates scale and channel attention to emphasize informative frequency bands and EMG electrodes. The Wavelet State-Space Model (WSSM) extends this framework with a gated state-space module that captures temporal dependencies in gesture dynamics.

The main contributions of this work are summarized as follows:

- A wavelet-based EMG gesture recognition framework that converts multi-channel EMG signals into CWT-based time–frequency representations to capture multi-scale muscle activation patterns.
- A Wavelet Attention Convolutional Neural Network (WA-CNN) that incorporates scale and channel attention mechanisms to emphasize informative frequency bands and EMG channels.
- A Wavelet State-Space Model (WSSM) that introduces efficient state-space temporal modeling for EMG gesture dynamics.
- A lightweight framework suitable for scalable training and potential deployment in wearable human–machine interaction systems.

The remainder of this paper is organized as follows. Section II reviews the related work on EMG-based gesture recognition. Section III describes the proposed wavelet-based framework, including the signal processing pipeline and the architectures of the WA-CNN and WSSM models. Section IV presents the experimental setup, performance evaluation, and comparison with baseline methods. Finally, Section V concludes the paper and discusses potential applications of the proposed approach.

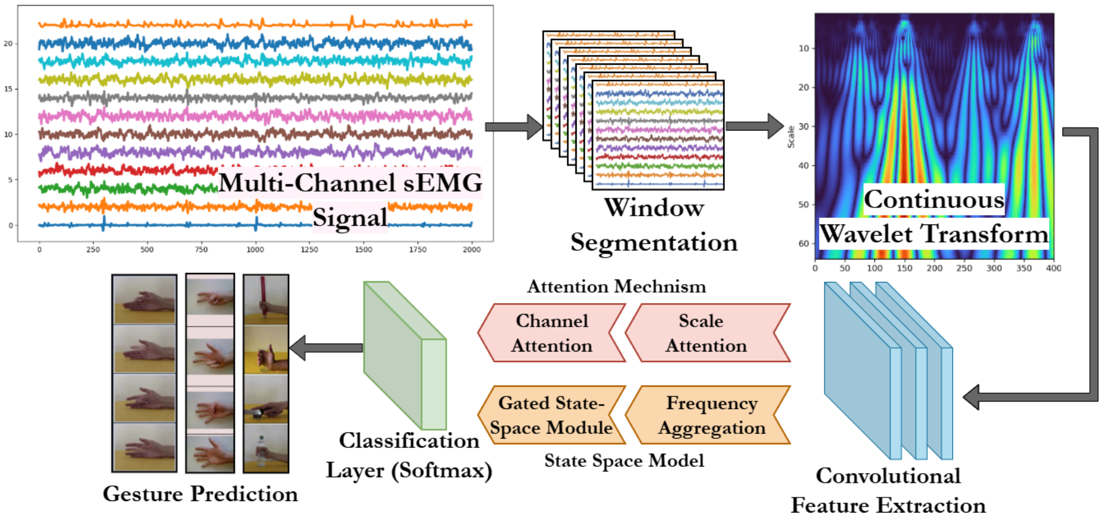

Fig. 1. Overview of the proposed wavelet-based EMG gesture recognition framework.

II. Related Work

Recent studies have increasingly explored deep learning techniques for surface electromyography (sEMG) based gesture recognition due to their ability to learn discriminative representations directly from multi-channel EMG signals. Rani *et al.* [7] employed convolutional neural networks to learn hierarchical representations from multi-channel EMG signals for gesture classification. Similarly, Rani *et al.* [8] introduced an SHAP-based deep learning framework to improve feature selection from EMG signals. Although these approaches improved recognition accuracy, many of them operated directly on raw EMG signals, which limited their ability to explicitly model the time–frequency structure of muscle activation patterns.

To address this limitation, several studies investigated time–frequency representations of EMG signals. Zhang *et al.* [9] transformed EMG signals into image-like representations using a 3D compression strategy and applied a broad learning system for gesture recognition. While this representation improved feature extraction, the method relied on shallow learning architectures that may limit the ability to capture complex hierarchical patterns. Similarly, Rani *et al.* [10] explored machine learning methods for EMG signal analysis and demonstrated improved classification performance using learned representations. However, the approach did not explicitly incorporate mechanisms for emphasizing informative frequency components of EMG signals.

Recent studies have also extended the investigation of more advanced sequence modeling techniques to effectively incorporate the temporal dependencies found in the EMG signal. In this context, Zabihi *et al.* [11] evaluated transformer-based architectures for modeling temporal dependencies in EMG gesture sequences. Even though the transformer models can improve the performance of the temporal modeling, the quadratic computational complexity may limit the real-time wearable systems.

Considering the above limitations, this study develops a framework that utilizes the EMG signal representation using Continuous Wavelet Transform, along with the efficient combination of attention-based convolutional learning with state space temporal modeling. The goal of this design is the efficient representation of discriminative time-frequency information, as well as maintaining efficiency in the context of wearable EMG-based gesture recognition systems.

III. Proposed Method

The following section presents a proposed wavelet-based method for EMG gesture recognition. The steps include EMG signal segmentation, performing time-frequency analysis using Continuous Wavelet Transform (CWT), and then using lightweight neural networks for classification. The framework is designed to effectively analyze EMG signals with intricate features while being computationally feasible for wearable human-machine interfaces. Table I presents the layer-wise configuration of the proposed architectures.

A. Dataset Description

The proposed framework is tested using the Ninapro DB2 dataset [12], a multichannel surface electromyography (sEMG) corpus that contains sEMG signals collected from individuals performing a wide range of gestures using their hands. Ninapro DB2 contains 40 gestures recorded from 40 subjects using 12 sEMG electrodes sampled at 2 kHz. The sEMG signals were collected using various electrodes positioned on the forearm to capture specific patterns of muscle activity associated with finger and hand gestures. The multichannel sEMG signals enable learning spatial relationships between various muscles.

B. Signal Segmentation

Raw EMG signals are segmented into fixed-length windows before feature extraction. Let $x_c(t)$ denote the EMG signal from channel c. The signal is divided into overlapping segments of length T samples using a sliding window approach. This segmentation converts the continuous EMG stream into a sequence of gesture-related signal segments:

$$X_c^{(i)} = x_c(t_i : t_i + T) \tag{1}$$

where $X_c^{(i)}$ represents the i-th segment of channel c. Window-based segmentation enables the model to learn localized muscle activation patterns while also increasing the number of training samples.

C. Continuous Wavelet Transform Representation

EMG signals are non-stationary and contain transient bursts of muscle activation. To capture these characteristics, each segmented signal is transformed into a time–frequency representation using the Continuous Wavelet Transform (CWT). The CWT of a signal $x(t)$ is defined as

$$W(a,b) = \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt \tag{2}$$

where a represents the scale parameter controlling frequency resolution, b denotes the time shift, and $\psi(\cdot)$ is the mother wavelet.

For multi-channel EMG signals, the CWT is computed independently for each channel. The resulting representations are stacked to form a multi-channel time–frequency tensor that serves as input to the neural network. This representation exposes structured patterns corresponding to different muscle activation frequencies and durations. The CWT transformation is computed on-the-fly during training to avoid storing large spectrogram datasets while preserving full time–frequency information during training and inference.

D. Wavelet Attention Convolutional Neural Network

The Wavelet Attention Convolutional Neural Network (WA-CNN) is designed to learn discriminative features from CWT spectrograms. The architecture uses stacked convolutional layers to extract hierarchical time–frequency patterns, where convolutional filters capture localized spectral and temporal correlations corresponding to muscle activation patterns.

To improve feature representation, two attention mechanisms are introduced. A scale attention module adaptively weights different wavelet scales, emphasizing informative frequency bands associated with specific gestures. In addition, a channel attention module highlights the most informative EMG electrodes while suppressing less relevant channels.

Let $F \in \mathbb{R}^{C \times S \times T}$ denote the feature tensor extracted by the convolutional backbone, where C, S, and T represent the channel, scale, and temporal dimensions, respectively. The attention-enhanced representation is computed as

$$F_a = (\alpha_c \otimes \alpha_s) \odot F \tag{3}$$

where $\alpha_c \in \mathbb{R}^C$ and $\alpha_s \in \mathbb{R}^S$ denote channel and scale attention weights, $\otimes$ represents the outer product used to expand the weights across dimensions, and $\odot$ denotes element-wise multiplication.

Algorithm 1: Training Procedure for Wavelet-Based EMG Gesture Recognition

Input: Multi-channel EMG signal $x_c(t)$, window length T, dataset $\mathcal{D}$
Output: Trained model parameters θ
Initialize network parameters θ;
for *each training epoch* **do**
  for *each EMG recording in* $\mathcal{D}$ **do**
    Segment signal using sliding window:

$$X_c^{(i)} = x_c(t_i : t_i + T)$$

    Compute Continuous Wavelet Transform:

$$W_c^{(i)}(a,b) = \int x_c^{(i)}(t) \psi^* \left(\frac{t-b}{a} \right) dt$$

    Stack wavelet coefficients to form input tensor

$$S^{(i)} \in \mathbb{R}^{C \times F \times T}$$

    Extract features using convolutional backbone:

$$F^{(i)} = \mathrm{CNN}(S^{(i)})$$

    if *model = WA-CNN* **then**
      Apply scale and channel attention

$$F_a^{(i)} = \mathrm{Attention}(F^{(i)})$$

    end
    if *model = WSSM* **then**
      Aggregate frequency features and update latent state

$$h_t = A h_{t-1} + B F^{(i)}$$

    end
    Compute class probabilities

$$\hat{y} = \mathrm{Softmax}(WF)$$

    Compute cross-entropy loss

$$\mathcal{L} = -\sum y \log(\hat{y})$$

    Update parameters θ using gradient descent;
  end
end
return trained model parameters θ;

E. Wavelet State-Space Model

Although convolutional networks effectively capture local time–frequency patterns, gesture dynamics also involve temporal dependencies across longer durations. To model these dependencies efficiently, the Wavelet State-Space Model (WSSM) extends the feature extraction framework with a gated state-space module.

TABLE I
LAYER-WISE CONFIGURATION OF THE PROPOSED MODELS

Layer	Configuration
Conv Block 1	Conv(32,3×3)+BN+ReLU
Conv Block 2	Conv(64,3×3)+BN+ReLU
Scale Attention	Global Pooling + FC
Channel Attention	Global Pooling + FC
SSM Layer	Hidden Dim=128
Classifier	FC + Softmax

In this architecture, convolutional layers first extract time–frequency features from the CWT representation. The feature maps are then aggregated along the frequency dimension to produce a temporal feature sequence. This sequence is processed by a gated state-space module that maintains a latent state representing the evolving muscle activation pattern over time. The state update integrates current observations with the previous latent state, enabling efficient modeling of temporal dependencies.

The state evolution in WSSM follows a gated linear state-space formulation

$$h_t = g_t \odot (Ah_{t-1}) + (1 - g_t) \odot (Bf_t) \tag{4}$$

where h_t denotes the latent state at time step t, f_t is the input feature vector, and g_t controls the contribution of previous state and current observation. State-space models were selected because they provide linear computational complexity with respect to sequence length while effectively capturing long-range dependencies.

F. Training Strategy

Both WA-CNN and WSSM are trained using supervised learning. The networks receive CWT-based EMG representations as input and output gesture class probabilities. Model parameters are optimized using the cross-entropy loss function defined as

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

where y_i represents the ground truth label and $\hat{y}_i$ denotes the predicted probability for class i. The training process enables the networks to learn discriminative time–frequency patterns associated with different hand gestures. Algorithm 1 summarizes the overall training pipeline. or subject-dependent evaluation, samples from each subject were randomly divided into 80% training and 20% testing sets. In addition, a cross-subject evaluation was conducted to assess the generalization capability of the proposed models on unseen subjects. Models are trained using Adam optimizer with a learning rate of 0.001 and batch size of 64 for 100 epochs and cross-entropy loss is employed.

IV. EXPERIMENTAL RESULTS

This section evaluates the effectiveness of the proposed wavelet-based EMG gesture recognition framework. The experiments compare the proposed models with conventional machine learning methods and deep learning architectures using multiple evaluation metrics.

A. Experimental Setup

Experiments were conducted on the Ninapro DB2 dataset. Performance was evaluated using classification accuracy and F1-score, while efficiency was assessed through parameter count and inference latency. All models were implemented in PyTorch and trained on an NVIDIA RTX 2070 GPU.

B. Comparison with Baseline Methods

Table II presents the performance comparison between conventional machine learning models, deep learning baselines, and the proposed architectures .

TABLE II
PERFORMANCE COMPARISON OF EMG GESTURE RECOGNITION MODELS

Model	Input	Acc(%)	F1	Params	Time
SVM	Handcrafted	55	0.54	0.45M	4 ms
Random Forest	Handcrafted	61	0.60	1.2M	6 ms
CNN (1D)	Raw EMG	76	0.75	11.7M	12 ms
CNN (2D)	STFT	87	0.86	8.0M	14 ms
CNN (2D)	CWT	89	0.88	22M	18 ms
Vision Transformer	CWT	93	0.54	22M	18 ms
ResNet18	CWT	91	0.60	11.7M	12 ms
WA-CNN (Proposed)	**CWT+Attention**	**92**	**0.91**	**2.3M**	**7 ms**
WSSM (Proposed)	**CWT+SSM**	**95**	**0.94**	**2.6M**	**8 ms**

Traditional machine learning models using handcrafted features achieve limited performance due to their inability to capture the complex temporal dynamics of EMG signals. Deep learning models trained on raw EMG improve performance but still fail to exploit frequency-dependent activation patterns fully. Using time–frequency representations significantly improve recognition accuracy. The proposed WA-CNN further enhances performance through scale and channel attention, while WSSM achieves the best results (95% accuracy) by modelling temporal dependencies using the state-space module.

C. Time–Frequency Representation Analysis

Fig. 2 illustrates the transformation of EMG signals from the raw waveform to time–frequency representations using STFT and CWT. The CWT representation provides clearer multi-scale activation patterns compared to STFT, which uses a fixed time-frequency resolution. This property enables the neural networks to capture transient muscle activation patterns more effectively.

D. Ablation Study

To analyze the contribution of each component of the proposed architecture, an ablation study is performed by progressively adding different modules to the baseline CNN model. The results are summarized in Table III.

The results demonstrate that the wavelet representation significantly improves performance compared to raw EMG

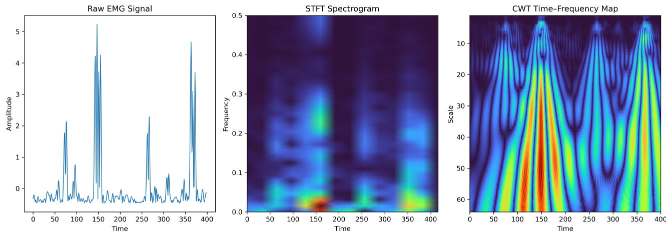

Fig. 2. Example EMG signal representation. From left to right: raw EMG waveform, STFT spectrogram, and CWT time–frequency map.

TABLE III
ABLATION STUDY OF THE PROPOSED ARCHITECTURE

Model Variant	Accuracy (%)
CNN (Raw EMG)	76
CNN + CWT	89
CNN + CWT + Channel Attention	90
CNN + CWT + Scale Attention	91
WA-CNN (Full Model)	92
WSSM	**95**

inputs. Channel attention and scale attention further enhance recognition accuracy by highlighting informative electrodes and frequency bands. The state-space temporal module provides the largest improvement by modeling gesture dynamics across time.

E. Cross-Subject Generalization

A cross-subject evaluation was conducted using a leave-subjects-out protocol. As shown in Table IV, performance decreases compared with the subject-dependent setting due to inter-subject variability. However, both WA-CNN and WSSM outperform the baseline CNN, with WSSM achieving the best cross-subject accuracy.

TABLE IV
CROSS-SUBJECT GENERALIZATION PERFORMANCE ON NINAPRO DB2

Model	Subject-Dependent	Cross-Subject
CNN (Raw EMG)	76	61
WA-CNN	92	86
WSSM	**95**	**89**

F. Computational Complexity

The computational cost of the proposed framework mainly arises from attention mechanism and the temporal state-space modeling. For conventional self-attention mechanisms, the pairwise interaction between all temporal tokens leads to a quadratic computational complexity given by $\mathcal{O}(T^2 \cdot d)$, since the attention matrix computation requires interactions across all temporal positions. Whereas for the WSSM model, the temporal modeling module follows a linear state-space formulation which results in a computational complexity of $\mathcal{O}(T \cdot d)$. Here, T denotes the sequence length and d denotes the latent state dimension. Unlike recurrent neural networks or transformer-based architectures that often exhibit $\mathcal{O}(T^2)$ complexity due to sequential or attention operations, the state-space formulation scales linearly with sequence length.

V. CONCLUSION

This work presents a wavelet-based deep learning framework for EMG gesture recognition using CWT time–frequency representations of sEMG signals. Two lightweight models, WA-CNN and WSSM, are developed to learn discriminative gesture patterns using attention mechanisms and state-space temporal modeling. Experimental results on the Ninapro DB2 dataset show that the proposed models outperform conventional machine learning and deep learning baselines while maintaining low computational complexity. In particular, WSSM achieves the best performance with 95% accuracy and efficient inference suitable for wearable systems. However, challenges remain in handling inter-subject variability and signal noise. Future work will address these through adaptive domain generalization and multimodal integration with vision and speech signals for robust human–machine interaction.

REFERENCES

[1] L. Wang, J. Fu, H. Chen, and B. Zheng, "Hand gesture recognition using smooth wavelet packet transformation and hybrid cnn based on surface emg and accelerometer signal," *Biomedical Signal Processing and Control*, vol. 86, p. 105141, 2023.

[2] H. Zhou, H. Thanh Le, S. Zhang, S. Lam Phung, and G. Alici, "Hand gesture recognition from surface electromyography signals with graph convolutional network and attention mechanisms," *IEEE Sensors Journal*, vol. 25, no. 5, pp. 9081–9092, 2025.

[3] J. Shin, A. S. M. Miah, S. Konnai *et al.*, "Hand gesture recognition using semg signals with a multi-stream time-varying feature enhancement approach," *Scientific Reports*, vol. 14, p. 22061, 2024.

[4] K. Challa, I. W. AlHmoud, A. K. M. Kamrul Islam, and B. Gokaraju, "Emg-based hand gesture recognition using individual sensors on different muscle groups," in *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2023, pp. 1–4.

[5] M. R. Islam, D. Massicotte, P. Massicotte, and W.-P. Zhu, "Surface emg-based intersession/intersubject gesture recognition by leveraging lightweight all-convnet and transfer learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–16, 2024.

[6] M. Shabanpour, K. Laamerad, S. Khademi, and A. Mohammadi, "Moemba: A mamba-based mixture of experts for high-density emg-based hand gesture recognition," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1–6.

[7] P. Rani, S. Pancholi, V. Shaw, M. Atzori, and S. Kumar, "Enhanced emg-based hand gesture classification in real-world scenarios: Mitigating dynamic factors with tempo-spatial wavelet transform and deep learning," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, no. 3, pp. 1202–1211, 2024.

[8] P. Rani, S. Pancholi, V. Shaw, S. Pandey, M. Atzori, and S. Kumar, "Explainable ai-guided optimization of emg channels and features for precise hand gesture classification: A shap-based study," *IEEE Transactions on Medical Robotics and Bionics*, vol. 7, no. 1, pp. 368–376, 2025.

[9] L. Zhang, H. Zhao, S. Qi *et al.*, "Improving semg signals recognition accuracy using 3d compression and broad learning system," *Signal, Image and Video Processing*, vol. 19, p. 1106, 2025.

[10] P. Rani, S. Pancholi, V. Shaw, M. Atzori, and S. Kumar, "Enhancing gesture classification using active emg band and advanced feature extraction technique," *IEEE Sensors Journal*, vol. 24, no. 4, pp. 5246–5255, 2024.

[11] S. Zabihi, E. Rahimian, A. Asif, and A. Mohammadi, "Trahgr: Transformer for hand gesture recognition via electromyography," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4211–4224, 2023.

[12] M. Atzori, A. Gijsberts, C. Castellini *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data*, vol. 1, p. 140053, 2014.