Attentional Convolutional Neural Networks Meet Kolmogorov-Arnold Networks for sEMG-based Gesture Recognition

Aiguo Wang*
School of Computer Science and Artificial Intelligence
Foshan University
Foshan, China
agwang@fosu.edu.cn

Sijie Wang
School of Computer Science and Artificial Intelligence
Foshan University
Foshan, China
heyyuan03@foxmail.com

Junjie He
School of Computer Science and Artificial Intelligence
Foshan University
Foshan, China
graversama@outlook.com

2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) | 979-8-3315-7423-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICAITA67588.2025.11137884

***Abstract*—Surface electromyography (sEMG) reflects muscle contraction states, provides critical signals for decoding hand movements, and thus holds significant potential in applications such as human-computer interaction and prosthetic control. However, it still a significant challenge to model the complex, non-linear, and dynamic patterns in the sEMG signals. In this paper, we propose a novel deep neural network that integrates convolutional neural network (CNN) and Kolmogorov-Arnold networks (KAN), termed ACNN-KAN, to utilize both CNN's local context modeling and KAN's adaptive nonlinearity for gesture recognition. Specifically, CNNs are first utilized to extract spatial features and an attention mechanism is integrated to enhance global dependency modeling. KAN is then leveraged to learn feature representation. We evaluate ACNN-KAN on three publicly available datasets (i.e., Ninapro DB1, DB5, and Myo) against nine competitors in terms of four performance metrics. Experimental results show that ACNN-KAN outperforms its competitors and achieves accuracies of 88.59% on DB1, 90.42% on DB5, and 96.10% on Myo, demonstrating its adaptability of KAN in gesture recognition.**

Keywords—Surface electromyography, gesture recognition, Kolmogorov-Arnold networks

I. INTRODUCTION

Gesture provides a natural and flexible means of conveying information in human-computer interaction, and hence gesture recognition has emerged as a prominent research area, with a variety of applications such as smart prosthetics, sign language recognition, and medical rehabilitation. However, the inherent complexity of hand movements poses a significant challenge for the accurate recognition of gestures. To enhance the accuracy and generalizability of gesture recognition systems, researchers have explored a wealth of sensing units and models. According to the underlying sensor types, existing methods can be divided into four groups: ambient sensor-based, vision-based, wearable sensor-based, and physiological sensor-based methods. Vision-based methods utilize cameras to capture image sequences of hand movements and use computer vision techniques to infer gestures [1]. Although having a wide range of applications, such methods are highly susceptible to lighting conditions, occlusions, and complex backgrounds, and inevitably raise privacy concerns. Ambient sensor-based methods detect hand movements by analyzing radar or WiFi signals [2]. Although offering a non-invasive solution, such methods suffer from low spatial resolution, which limits their ability to detect fine-grained hand movements and makes them prone to interference from other signals. Wearable sensor-based methods exploit devices such as accelerometers and gyroscopes to track hand movements [3]. Though having the advantage of pervasiveness, their performance is easily influenced by sensor placement variability and interference from body movements. Since hand movements elicit distinct physiological responses, researchers have also investigated sEMG-based approaches due to their high sensitivity, non-invasive nature, and less susceptibility to environmental interference [4]. sEMG signals provide valuable insights into muscle activity and serve as key indicators of hand movements. Traditionally, the sEMG-based gesture recognition chain consists of the training stage and prediction stage. For the former, a gesture recognition model is trained on the collected sensor data, where the extraction of features and choice of classification models are its two crucial components. Feature extraction aims to extract handcrafted meaningful representations from raw sEMG signals to better reflect the characteristics of gestures. We can group existing features into time domain, frequency domain, and time-frequency domain. Commonly used time-domain features include mean, maximum, minimum, and zero cross rate. The fast Fourier transform is used to transform time-domain signals into frequency domain, followed by the extraction of features such as energy, direct component, spectral centroid, skewness, and kurtosis. Wavelet transform and Hilbert-Huang transform can be used to extract time-frequency features. For example, Zhang et al. proposed a gesture recognition model trained with time-domain features to distinguish five gestures [5]. Afterwards, the classification model takes as input the extracted features to optimize a gesture recognizer. The candidate classification models range from discriminative to generative models. One major limitation of traditional machine learning based methods is that their performance largely relies on the quality and quantity of the selected features [6].

The rapid advancement of deep learning has significantly accelerated research in gesture recognition. Deep learning techniques enable end-to-end learning by automatically extracting hierarchical representations from raw sEMG data, which eliminates the need for manual feature engineering and obtains superior performance. To utilize translation and rotation invariance among sEMG signals, researchers have explored the

Authorized licensed use limited to: Tsinghua University. Downloaded on September 10,2025 at 02:39:19 UTC from IEEE Xplore. Restrictions apply.

use of convolutional neural networks. For example, Liu et al. designed a multiscale CNN-based gesture recognition model for recognizing sixteen complex hand gestures [7]. To capture the long-range dependency among signals, researchers have also investigated the power of recurrent neural networks. For example, He et al. utilized long short-term memory (LSTM) networks to recognize static gesture features, obtaining accuracy of 75.45% for ten gestures on the Ninapro dataset [8]. Recognizing the critical role of attention mechanisms in dynamically highlighting relevant features, recent studies have integrated these mechanisms into gesture recognition models. For example, Xu et al. designed a CNN model augmented with channel-wise attention and obtained accuracy of 89.54% for 18 gestures on the Ninapro DB5 dataset [9]. Despite improvements in recognition accuracy, these methods usually require substantial computational resources and face two challenges: the locality of convolution operations constrains the ability of CNNs to capture long-term dependencies, which is crucial for decoding gestures with complex temporal dynamics; the fully connected layers rely on static linear transformations and are sensitive to parameter tuning, where fixed activation functions struggle to adapt to the non-stationary characteristics of sEMG signals. To overcome these limitations, the Kolmogorov-Arnold Network (KAN), grounded by the Kolmogorov-Arnold representation theorem, has been proposed. KAN effectively captures complex dependencies and patterns within data, making it particularly well-suited for gesture recognition. Furthermore, its integration with CNNs can further enhance recognition accuracy. By leveraging the local context modeling capability of CNNs and the adaptive nonlinearity of KAN, this hybrid architecture achieves a balance between flexibility and computational efficiency. In this study, we propose a deep neural network that integrates CNNs and KANs, termed ACNN-KAN, to enhance feature extraction and representation. Specifically, CNNs are utilized to extract spatial features, while an attention mechanism is incorporated to capture global dependency. KAN is then utilized to refine feature representation, where learnable activation functions are employed to enhance the expressiveness of fully connected layers. The main contributions of this study are as follows. (1) We propose a hybrid architecture of CNNs and KAN. This integration enables the model to utilize CNN's local context modeling ability and KAN's adaptive nonlinearity. Particularly, an attention mechanism is embedded in CNNs to selectively emphasize salient features and suppress irrelevant ones. (2) We conduct comparative experiments on three public datasets to validate the effectiveness of ACNN-KAN. Comparative experiments against nine competitors in terms of four performance metrics demonstrate its superiority.

II. GESTURE RECOGNITION MODEL

A. The Proposed Model

Fig. 1 presents the model architecture of ACNN-KAN that consists of three main components: basic CNN architecture, Squeeze-and-Excitation (SE) block, and KANLinear layer. The basic CNN architecture comprises four Conv2D layers, each setting the kernel size to 3x3 and using PReLU as the activation function. Batch normalization follows each Conv2D layer to facilitate training stability. To enhance the network's feature representation capability, one SE block is inserted after each PReLU activation function. Afterwards, KANLinear layer is used to learn features and perform feature mapping, where KANLinear layer adopts learnable activation functions for its weight components.

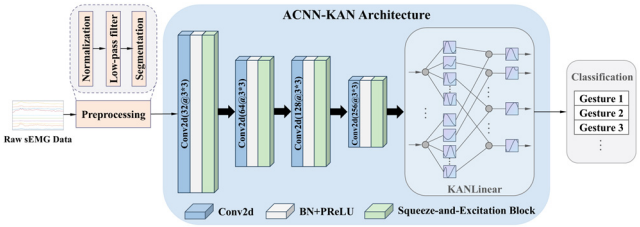

Fig. 1. Schematic diagram of the network structure of ACNN-KAN.

B. Kolmogorov-Arnold Network

The Kolmogorov-Arnold representation theorem indicates that any multivariate continuous function f can be represented as a combination of a finite number of univariate continuous functions: $f(x_1,\cdots,x_n)=\sum_{q=1}^{2n+1}\Phi_q\left(\sum_{p=1}^{n}\varphi_{q,p}(x_p)\right)$, where $\varphi_{q,p}$ is the univariate function that maps each input variable $x_p \in [0,1]$ to $\mathbb{R}$, and $\Phi_q:\mathbb{R}\to\mathbb{R}$. With this theorem, the Kolmogorov-Arnold Network is proposed [10], where the KANLinear layer is used to build a neural network layer. KANLinear employs learnable univariate functions to approximate complex multivariate functions $y=\sum_{q=1}^{Q}\Phi_q\left(\sum_{p=1}^{P}\varphi_{q,p}\left(x_p\right)\right)$, where x_p is the p^{th} input feature, with P being the total number of input features, Q is the number of output features, and $\varphi_{q,p}$ and Φ_q are univariate functions realized via splines, respectively. The function spline(x) is parameterized as a linear combination of B-spline.

$$\text{spline}(x)=\sum_i c_i B_i(x) \tag{1}$$

, where $B_i(x)$ is the B-spline basis functions and c_i are the trainable coefficients. These coefficients determine the activation function. To further enhance the stability and trainability, the KANLinear layer introduces a residual activation function $\varphi(x)$.

$$\varphi(x)=w_b b(x)+w_s\,\text{spline}(x) \tag{2}$$

, where $b(x)$ is the basis function. The residual structure provides a direct path for gradient propagation. Unlike traditional neural networks with fixed activation functions, KAN employs learnable activation functions, which allows the network to flexibly capture complex data patterns. It is worth noting that the original design of KAN may lead to longer training times. Hence, we use an optimized implementation of the original KAN, called Efficient KAN [11].

III. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Data

We in this study use three public datasets (i.e., Ninapro DB1, DB5, and Myo [12]) as shown in Table 1 to evaluate the effectiveness of ACNN-KAN. DB1 comprises sensor data recorded from 27 able-bodied subjects using 10 OttoBock sEMG electrodes at a sampling rate of 100 Hz, with each gesture repeated 10 times. DB5 consists of data from 10 able-bodied

Authorized licensed use limited to: Tsinghua University. Downloaded on September 10,2025 at 02:39:19 UTC from IEEE Xplore. Restrictions apply.

subjects, recorded using two Thalmic Myo armbands at a sampling rate of 200 Hz, with each gesture repeated 6 times. Each of the datasets includes 52 movements (plus a rest position), categorized into Exercise *A*, *B*, and *C*. Our study focuses on Exercise *B* that consists of eight isometric and isotonic hand configurations, nine fundamental wrist movements, and a 'relax' gesture. The Myo dataset consists of data from 13 able-bodied subjects using a single Myo armband. The dataset consists of 21 gestures, with each gesture repeated 30 times.

Table 1 Basic Information of Experimental Datasets

Datasets	#Subjects	#Gestures	sEMG Channels	Sampling Rate
Ninapro DB1	27	18	10	100Hz
Ninapro DB5	10	18	16	200Hz
Myo	13	21	8	200Hz

B. Experimental Setup

Since raw sEMG signals are highly sensitive to noise and external interference and contain a significant amount of invalid data, preprocessing is performed to provide high-quality input for subsequent tasks. Specifically, data are first normalized to the range [0,1]. A fourth-order Butterworth low-pass filter with a 50 Hz cutoff frequency is applied to remove low-frequency noise. Afterwards, a sliding window of 1000 milliseconds with a stride of 10 milliseconds is used to segment the raw sEMG signals. To ensure an unbiased evaluation, we follow the official experimental setup for the Ninapro dataset. For DB5, the 2nd and 5th repetitions are designated as the test set; for DB1, the 2nd, 5th, and 7th repetitions are used as the test set. The remaining serve as the training and validation sets, where a further k-fold cross validation is used to generate independent training and validation sets. This procedure is repeated four times for DB5 (and seven times for DB1), and we report the average accuracy, precision, recall, and F1-score. For the Myo dataset, each gesture is repeated 30 times, which are split into training, validation, and test sets. In each fold, five repetitions are assigned as the test set, five as the validation set, and the remaining 25 for training. This process is repeated six times. The results are reported as the average accuracy, precision, recall, and F1-score.

To validate the effectiveness of the proposed model, we compared its performance against commonly used machine learning based methods (including support vector machine (SVM), random forest (RF)) and deep learning architectures (including CNN, LSTM, CNN-LSTM, temporal convolutional network (TCN), SE-CNN [9], KAN [11], and DRCN [13]). The features for traditional machine learning based methods include mean absolute value, slope sign change, waveform length, and root mean square. For deep learning-based models, they take as input the raw sEMG signals. Table 2 presents an overview of the architectural configurations of the compared models. To ensure a fair comparison, the models are trained using consistent strategies. We implement the models in the PyTorch framework and use the Adam optimizer for model training on a server equipped with an NVIDIA GeForce RTX 4090 GPU and an Intel® Core™ i7-13700KF 3.42 GHz CPU. We empirically set the initial learning rate of 0.001 and a batch size of 128. During training, if the validation loss does not decrease for 8 consecutive epochs, we reduce the learning rate by a factor of 10. The number of epochs is 100 and the training process is terminated if the validation loss does not improve for 10 consecutive epochs.

Table 2 Architectural Parameters of Different Deep Learning Models

Models	Modules	Parameters	Value
CNN	/	Layers	4
		Kernel size	[3, 3]
		Kernels	[32, 64, 128, 256]
LSTM	/	Layers	3
		Hidden nodes	128
CNN-LSTM	CNN	Layers	3
		Kernel size	3
		Kernels	[64, 128, 256]
	LSTM	Layers	3
		Hidden nodes	128
TCN	/	Levels	[64(d=1), 64(d=2), 128(d=4), 128(d=8)]
		Kernel size	2
		Hidden dimensions	[64, 64, 128, 128]
KAN	/	Grid size	5
		Spline order	3
		Scale noise	0.1
		Activation function	SiLU
		Grid range	[-1, 1]
SE-CNN	CNN	Layers	3
		Kernel size	[8, 5, 3]
		Kernels	[128, 256, 512]
	Squeeze-and-Excite block	Layers	3
	Attention	Layers	1
	/	Dropout	0.5
DRCN	CNN	Layers	1
		Kernel size	[7, 7]
		Kernels	16
	DRCN block	Layers	3
		Dilation	[1, 2, 4]
		Kernel size	[3, 3]
		Kernels	[32, 64, 128]
ACNN-KAN	CNN	Layers	4
		Kernel size	[3, 3]
		Kernels	[32, 64, 128, 256]
	Squeeze-and-Excite block	Layers	4
	KANLinear	Grid size	5
		Spline order	3
		Scale noise	0.1
		Activation function	SiLU
		Grid range	[-1, 1]

C. Results

Table 3 presents the results concerning accuracy, precision, recall, and F1 score. We can observe that ACNN-KAN model outperforms its competitors across the three datasets. For example, ACNN-KAN achieves an accuracy of 88.59%, 90.42%, and 96.10% on DB1, DB5, and Myo, respectively, surpassing the best-performing DRCN model by a margin of 0.74%, 0.91%, and 0.81%. We can also observe that machine learning-based methods can be competitive in some scenarios. For example, SVM and RF on DB1 achieve accuracies of 84.18% and 84.44%, respectively, outperforming both TCN (73.62%) and LSTM (73.91%). However, these methods struggle to model the complex spatiotemporal features of raw sEMG signals. For instance, on the Myo dataset, the accuracy of SVM (90.33%) is notably lower than that of ACNN-KAN (96.10%). Among the

Authorized licensed use limited to: Tsinghua University. Downloaded on September 10,2025 at 02:39:19 UTC from IEEE Xplore. Restrictions apply.

four deep learning models (i.e., CNN, LSTM, CNN-LSTM, and TCN), CNN consistently outperforms the other three across all datasets. For example, CNN on DB1 achieves 86.64% accuracy, outperforming LSTM (73.91%), CNN-LSTM (84.37%), and TCN (73.62%). This is due to CNN's capability to capture local and translation invariance in sEMG signals. KAN performs better than TCN and LSTM across all datasets, which can be attributed to the replacement of traditional linear weight matrices with learnable activation functions of KAN. This modification enables more effective feature learning. Integrating KAN's robust function approximation capability with CNN's local feature extraction ability leads to classification performance gains. For example, ACNN-KAN on DB5 surpasses KAN by 4.85% and CNN by 3.66% in accuracy. Besides, we also compared it with other sEMG-based gesture recognition methods reported in the literature (i.e., SE-CNN and DRCN). The results confirm that ACNN-KAN outperforms these models across all evaluated datasets.

Table 3 Results of Different Gesture Recognition Models

Models	**Ninapro DB1**				**Ninapro DB5**				**Myo**			
	Accuracy	***Precision***	***Recall***	***F1***	***Accuracy***	***Precision***	***Recall***	***F1***	***Accuracy***	***Precision***	***Recall***	***F1***
SVM	0.8418	0.8482	0.8430	0.8398	0.7940	0.8013	0.7961	0.7923	0.9033	0.9110	0.9033	0.9019
RF	0.8444	0.8497	0.8458	0.8419	0.8467	0.8534	0.8488	0.8457	0.8985	0.9062	0.8985	0.8965
CNN	0.8664	0.8802	0.8686	0.8660	0.8676	0.8882	0.8717	0.8695	0.9524	0.9579	0.9524	0.9511
LSTM	0.7391	0.7454	0.7388	0.7314	0.6124	0.6053	0.6039	0.5889	0.8914	0.8988	0.8914	0.8887
CNN-LSTM	0.8437	0.8554	0.8461	0.8421	0.8211	0.8320	0.8220	0.8165	0.9190	0.9262	0.9190	0.9175
TCN	0.7362	0.7499	0.7379	0.7349	0.6735	0.6965	0.6678	0.6607	0.6993	0.7074	0.6993	0.6954
KAN	0.8185	0.8345	0.8202	0.8182	0.8557	0.8678	0.8571	0.8545	0.9100	0.9161	0.9100	0.9079
SE-CNN	0.8577	0.8706	0.8605	0.8557	0.8321	0.8451	0.8319	0.8277	0.9215	0.9286	0.9215	0.9191
DRCN	0.8785	0.8882	0.8799	0.8776	0.8951	0.9043	0.8975	0.8950	0.9529	0.9577	0.9530	0.9516
ours	0.8859	0.8934	0.8866	0.8842	0.9042	0.9153	0.9059	0.9045	0.9610	0.9650	0.9610	0.9599

IV. CONCLUSION

Due to the inherent complexity of gestures, how to capture non-linear dependencies and learn feature representations largely determines the performance of a gesture recognition model. In this paper, we propose a novel sEMG-based gesture recognition model that integrates convolutional neural networks and Kolmogorov-Arnold networks to enhance classification accuracy. By leveraging CNNs for spatial feature extraction and an attention mechanism to highlight discriminative regions, our model effectively captures critical muscle activity patterns. Besides, the use of KAN enables better handling of non-linear feature representations. Finally, comparative experiments on three datasets demonstrate the superiority of our approach.

For future work, deep learning-based gesture recognition models generally have high computational requirements. To improve the general applicability of these models, we plan to use lightweight techniques to reduce model complexity. Second, due to individual differences, cross-domain performance of gesture recognition models suffers from degraded performance. Therefore, leveraging transfer learning techniques to enhance domain adaptability remains another promising research topic.

REFERENCES

[1] A. Wang, H. Liu, C. Zheng, H. Chen, and C. Y. Chang, "Fusion of kinematic and physiological sensors for hand gesture recognition," Multimed. Tools Appl., vol. 83, no. 26, pp. 68013-68040, January 2024.

[2] M. Yang, H. Zhu, R. Zhu, F. Wu, L. Yin, and Y. Yang, "WiTransformer: A novel robust gesture recognition sensing model with WiFi," Sensors, vol. 23, no. 5, pp. 2612, February 2023.

[3] S. Zhang, H. Zhou, R. Tchantchane, and G. Alici, "Hand gesture recognition across various limb positions using a multi-modal sensing system based on self-adaptive data-fusion and convolutional neural networks (CNNs)," IEEE Sensors J., vol. 24, no. 11, pp. 18633-18645, June 2024.

[4] A. Wang, H. Liu, and J. Yan, "Hand gesture recognition using multi-sensor information fusion," In Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), pp. 153, April 2023.

[5] Z. Zhang, K. Yang, J. Qian, and L. Zhang, "Real-time surface EMG pattern recognition for hand gestures based on an artificial neural network," Sensors, vol. 19, no. 14, pp. 3170, July 2019.

[6] M. Rahman, A. Uzzaman, F. Khatun, M. Aktaruzzaman, and N. Siddique, "A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems," Expert Syst. Appl., vol. 266, p. 125929, December 2024.

[7] T. Liu, D. Bai, L. Ma, Q. Du, and H. Yokoi, "Complex surface electromyography signal gesture recognition based on multi-pathway featured scale convolutional neural network," IEEE Trans. Instrum. Meas., vol. 73, 2535311, October 2024.

[8] Y. He, O. Fukuda, N. Bu, H. Okumura, and N. Yamaguchi, "Surface EMG pattern recognition using long short-term memory combined with multilayer perceptron," In 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5636-5639, July 2018.

[9] Z. Xu, J. Yu, W. Xiang, S. Zhu, B. Liu, and J. Li, "A novel SE-CNN attention architecture for sEMG-based hand gesture recognition," CMES-Comp. Model. Eng., vol. 134, no. 1, pp. 157-177, January 2023.

[10] Z. Liu, Y Wang, S Vaidya, et al, "Kan: Kolmogorov-arnold networks," 2024, *arXiv:2404.19756*.

[11] Blealtan, "Efficient-KAN: An efficient pure-pytorch implementation of kolmogorov-arnold network (KAN)," GitHub repository, 2024. [Online]. Available: https://github.com/Blealtan/efficient-kan

[12] Z. Zhang, C. He, and K. Yang, "A novel surface electromyographic signal-based hand gesture prediction using a recurrent neural network," Sensors, vol. 20, no. 14, pp. 3994, July 2020.

[13] Z. Zhang, B. Zhao, X. Zhang, and Y. Zhang, "Dilated residual convolutional network for surface electromyographic hand gesture recognition," Biomed. Signal Process. Control, vol. 103, pp. 107438, May 2025.

Authorized licensed use limited to: Tsinghua University. Downloaded on September 10,2025 at 02:39:19 UTC from IEEE Xplore. Restrictions apply.