ViT-HGR: Vision Transformer-based Hand Gesture Recognition from High Density Surface EMG Signals*

Mansooreh Montazerin[1], Soheil Zabihi[1], Elahe Rahimian[2], Arash Mohammadi[1,2], and Farnoosh Naderkhani[2]

Abstract— Recently, there has been a surge of significant interest on application of Deep Learning (DL) models to autonomously perform hand gesture recognition using surface Electromyogram (sEMG) signals. DL models are, however, mainly designed to be applied on sparse sEMG signals. Furthermore, due to their complex structure, typically, we are faced with memory constraints; require large training times and a large number of training samples, and; there is the need to resort to data augmentation and/or transfer learning. In this paper, for the first time (to the best of our knowledge), we investigate and design a Vision Transformer (ViT) based architecture to perform hand gesture recognition from High Density (HD-sEMG) signals. Intuitively speaking, we capitalize on the recent breakthrough role of the transformer architecture in tackling different complex problems together with its potential for employing more input parallelization via its attention mechanism. The proposed Vision Transformer-based Hand Gesture Recognition (ViT-HGR) framework can overcome the aforementioned training time problems and can accurately classify a large number of hand gestures from scratch without any need for data augmentation and/or transfer learning. The efficiency of the proposed ViT-HGR framework is evaluated using a recently-released HD-sEMG dataset consisting of 65 isometric hand gestures. Our experiments with 64-sample ($31.25\ ms$) window size yield average test accuracy of $84.62 \pm 3.07\%$, where only $78,210$ number of parameters is utilized. The compact structure of the proposed ViT-based ViT-HGR framework (i.e., having significantly reduced number of trainable parameters) shows great potentials for its practical application for prosthetic control.

I. INTRODUCTION

Thanks to the recent evolution in the field of Artificial Intelligence (AI), specifically Deep Neural Networks (DNNs), significant advancements are expected on development of highly functional hand prostheses for upper limb amputees. Generally speaking, such advanced prosthesis systems are, typically, designed using surface Electromyogram (sEMG) signals [1]–[5], representing action potentials of the muscle fibers [6]. The sEMG signals, after passing through a pre-processing stage, could be a valuable input for DNN architectures to perform different tasks including but not limited to motor control, prosthetic device control, and/or hand motion classification. Researchers are, therefore, turning their attention to development of DL-based Human Machine Interface (HMI) algorithms using sEMG signals to design more accurate and more efficient myoelectric prosthesis control

systems. The ultimate goal is improving quality of life of individuals suffering from amputated limbs.

The sEMG signals are generally classified into two main categories, i.e., sparse and high-density [7]–[9]. Both of these multi-channel signals are acquired from a grid of electrodes placed on stump muscles to record the electrical activities and temporal information of the muscle tissue. Although using a large number of electrodes makes the computational process challenging, there has been a surge of recent interest in the use of High-Density sEMG (HD-sEMG) signals. More specifically, HD-sEMG signals are capable of recording both temporal and spatial information of the whole muscle [10]. Furthermore, by placing a large number of electrodes on the recording area, HD-sEMG provides noticeably high resolution signals compared to those obtained from sparse electrodes. This particular method improves the sparsity of the information that is collected from the muscle fibers, which results in improved classification efficiency. Moreover, these signals are independent of the exact position of the electrodes [11], [12], therefore, small variations in the position of the 2-Dimensional (2D) grid does not considerably impact the quality of the signals. Thus, development of advanced processing/learning models based on HD-sEMG signals is of significant importance to perform hand gesture recognition with high accuracy for design of myoelectric prosthesis.

Literature Review: Among the most prevalent approaches adopted for the task of hand gesture recognition, we can refer to the traditional Machine Learning (ML) methods such as Linear Discriminator Analysis (LDA) and Support Vector Machines (SVMs). These methods can achieve reasonable accuracy only when dealing with small data sets with a limited number of classes. However, when it comes to large data sets, most of conventional ML algorithms fail to perform at the same performance level. For such scenarios, DNNs are attractive alternatives to perform the recognition tasks with high accuracy. Convolutional Neural Networks (CNN) [13]–[16], for instance, have been widely used for classification of sEMG signals by converting them into 2D images. However, such an approach is not always practical because it only considers the spatial domain of sEMG signals, ignoring their sequential nature. To jointly incorporate temporal and spatial characteristics of multi-channel sEMG signals, several hybrid models, that combine CNNs with Recurrent Neural Networks (RNN), have been proposed [17], [18]. For instance, authors in [17] translated raw sEMG signals using six sEMG image representation techniques, which are then provided as input

[1] Department of Electrical and Computer Engineering, Concordat University, Montreal, Canada

[1]Concordia Institute for Information System Engineering (CITIES), Concordat University, Montreal, Canada

to a hybrid CNN-RNN model. The outcomes of [17] eventually show that the accuracy of image classification is highly dependent on the quality of the constructed image, making the key question on "*What would be the best procedure for producing images from sEMG signals?*" still not fully resolved [19]. In spite of their remarkable contributions to sequence modelling, hybrid CNN-RNN models fail to pass the entire computations for training to parallel blocks because of their innately sequential structure [20]. This raises some issues such as memory constraints and longer training times when working with large sequences. The paper aims to address this issue via incorporation of transformer architectures.

Contributions: In this paper, for the first time (to the best of our knowledge), we investigate and design a Transformer-based architecture [20] to perform hand gesture recognition from HD-sEMG signals. Intuitively speaking, we capitalize on the recent breakthrough role of the Transformer architecture in tackling different complex ML problems together with its great potential for employing more input parallelization with attention mechanism. Since HD-sEMG data sets have a 3-Dimensional (3D) structure, Vision Transformers (ViT) [21] can be considered as an appropriate architecture to address the challenges identified above. The proposed Vision Transformer-based Hand Gesture Recognition (ViT-HGR) framework can overcome training time problems and evaluation accuracy that we mostly face while working with other similar networks such as the conventional ML algorithms or the more advanced DNNs such as Long Short-Term Memories (LSTMs). However, we cannot directly provide HD-sEMG signals as input to the ViTs, and particular signal processing steps are required to modify the signal into a format that is compatible with the ViT's input. Therefore, the proposed ViT-HGR architecture converts the main signal into smaller portions using a specific window size and then feeds each of these portions to the ViT for further analysis. In brief, contributions of this work are as follows:

- A Transformer-based architecture, referred to as the ViT-HGR framework, is designed for the first time, to the best of our knowledge, to perform hand gesture recognition from HD-sEMG signals.
- As direct application of ViT to HD-sEMG is not possible and straightforward, a particular signal processing step is developed to convert the HD-sEMG signals to a specific format that is compatible with ViTs. In other words, the proposed ViT-based ViT-HGR framework can learn from HD-sEMG signals rather than images.
- The proposed ViT-HGR framework can accurately classify a large number of hand gestures from scratch without any need for data augmentation and/or transfer learning.

To develop and evaluate the proposed ViT-HGR framework, we used a recently-released HD-sEMG dataset consisting of 65 isometric hand gestures and 128 distinct channels for recording the signals [22]. Our results show superior performance of the ViT-HGR framework compared to its counterparts illustrating reduced training time and increased testing accuracy.

The remainder of the paper is organized as follows: The utilized dataset and initial preprocessing procedure are introduced in Section II. The proposed ViT-HGR framework together with steps used to prepare HD-sEMG signals to be fed to the ViT architecture are discussed in Section III. Experimental results are presented in Section IV. Finally, Section V concludes the paper.

II. MATERIALS AND METHODS

A. The Dataset

The proposed ViT-HGR framework is developed based on a database of HD-sEMG signals obtained from 20 subjects that performed 65 isometric hand gestures plus a gesture that was repeated twice (66 gestures in total) [22]. Following initial evaluations, one subjects is disregarded because of its partial information. Two grids each comprising of 64 electrodes (8×8) were placed on the flexor and extensor muscles of the participants and they were asked to repeat each movement 5 times with five seconds rest intervals. Signals were recorded via the Quattrocento (OT Bioelettronica, Torino, Italia) biomedical amplifier system and were sampled at $2,048$ Hz frequency. The train and test sets were not mentioned in [22]. Therefore, we perform a 5-fold cross validation to provide the bases for fair evaluations with future works. More specifically, the dataset is presented in five repetitions, where each time one repetition is used as the test set while the remaining repetitions are used as the training set. The test repetition, therefore, is changing each time and the final result is presented based on the average accuracy for each fold.

B. Initial Preprocessing

The raw dataset consists of plenty of sharp fluctuations that commonly occur in the EMG signals. Not only are these fluctuations required for an accurate gesture recognition task but they also prevent the network from training the useful information and increasing its accuracy. As a result, we filter the signals of each electrode separately with a first-order low-pass butterworth filter, which yields the positive envelope of each channel. Filtered signals are then passed through a μ-law normalization given by

$$F(x_t) = sign(x_t)\frac{\ln\left(1 + \mu|x_t|\right)}{\ln\left(1 + \mu\right)}, \quad (1)$$

where x_t represents the EMG signal for each channel and μ is the parameter to which the signal is scaled. As shown in [23], normalizing the signals increases the discriminative power of the network.

III. THE PROPOSED ViT-HGR FRAMEWORK

The proposed ViT-HGR framework is implemented based on the transformers and attention mechanism [21]. The attention mechanism incorporated with CNNs and the LSTMs were formerly used for the hand movement classification tasks because of their proven ability to leverage the temporal information of the sEMG signals [17]. However, in this work,
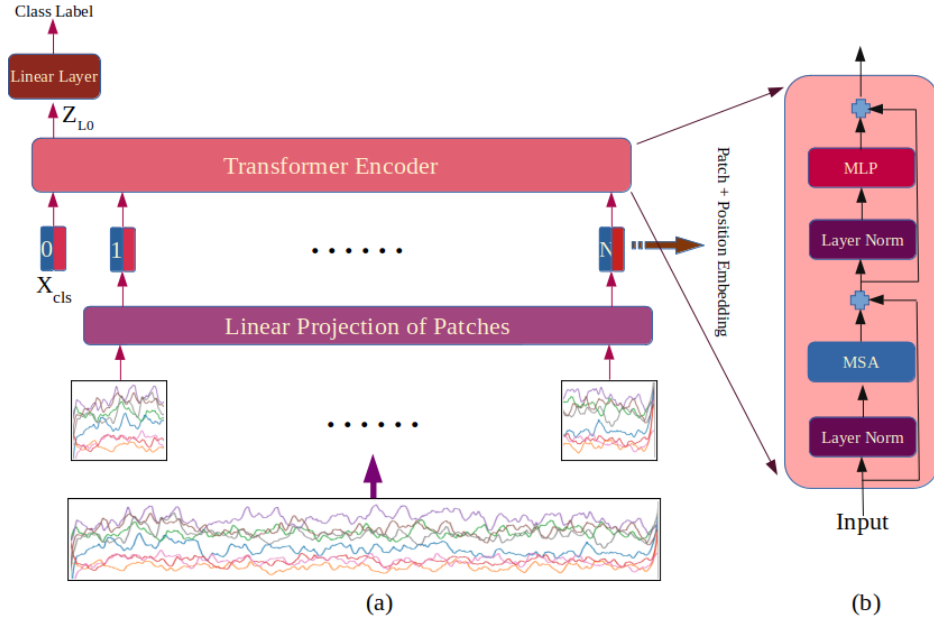
Fig. 1. A graphical representation of the Vision Transformer: (a) A cropped portion of the original signal is fed to the ViT and converted to small patches. These patches go through a patch + position embedding layer and a class token is prepended to them. They are then inputted to the transformer encoder. (b) The transformer encoder that is a combination of Multi-head Self-Attention, Multilayer Perceptron and two Add&Norm layers.

we indicate that the attention framework itself is sufficient to surpass the other networks and because of our data preparation approach, there is no need for data augmentation.

The overall structure of the proposed ViT-HGR architecture is shown in Fig. 1. The ViT is designed to have a 3D image as an input, which is then reshaped to a sequence of flat 2D patches with a predetermined length. In order to convert the HD-EMG signals to the acceptable input for the ViT, we first split the whole sequence into multiple sections utilizing a sliding window of 64 samples (31.25 ms considering the $2,048$ Hz sampling frequency) with a skip_step equal to 32 samples. It is worth mentioning that the window length should be less than 300 ms to satisfy the acceptable delay time [24], which is the real-time response required for practical myoelectric prosthetic control. Therefore, the window length for the classification purpose cannot surpass 300 ms [24]. Although using a larger window size potentially leads to better performance, use of shorter windows (e.g., 31.25 ms) is preferred allowing extra necessary time for practical implementation in real scenarios. After the windowing step, each of the (window_size, N_{ch}, N_{cv}) sequences, where N_{ch} is the number of horizontal channels and N_{cv} is the number of vertical channels, are considered as the 3D input of the ViT and is divided into small square patches in the Patch Embedding block. The sequence of patches is then fed to a linear projection trainable layer ($\boldsymbol{E}$) and a class token ($\mathbf{x}_{cls}$) is created and put at the beginning of this sequence. The class token here is exclusively responsible for holding all the learned information during the training and is then used in the last layer to detect the gestures. In order to maintain the positional information of the signals throughout the training phase, a positional embedding vector goes through the transformer's encoder block together with each patch. The aforementioned steps provides the input ($\boldsymbol{Z}_0$)

to the transformer encoder as follows:

$$\boldsymbol{Z}_0 = [\mathbf{x}_{cls}; \mathbf{x}_1^p \boldsymbol{E}; \mathbf{x}_2^p \boldsymbol{E}; \ldots; \mathbf{x}_N^p \boldsymbol{E}] + \boldsymbol{E}^{pos}. \tag{2}$$

The encoder block is a standard transformer encoder which is introduced by [20] and has L identical layers in total. Each layer consists of two principal blocks called the "Multi-Head Self Attention" (MSA) and the "Feed Forward". The former is assigned for the attention mechanism while the latter acts as a position-wise Multilayer Perceptron (MLP). In the MSA block, there are H number of identical heads with distinct learnable parameters operating in parallel. These heads separately accept $1/H$ of the total length of Keys ($\boldsymbol{K}$), Queries ($\boldsymbol{Q}$) and Values ($\boldsymbol{V}$) and after performing the following scaled dot-product attention

$$SA(\boldsymbol{Z}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_h}})\boldsymbol{V}, \tag{3}$$

$$MSA(\boldsymbol{Z}) = [SA_1(\boldsymbol{Z}); SA_2(\boldsymbol{Z}); \ldots; SA_h(\boldsymbol{Z})]W^{MSA}, \tag{4}$$

concatenate the outputs and transfer them to the MLP layer. The final output of the transformer after passing the MSA and MLP layers can be shown as the following matrix

$$\boldsymbol{Z}_L = [\boldsymbol{z}_{L0}; \boldsymbol{z}_{L1}; \ldots; \boldsymbol{z}_{LN}], \tag{5}$$

where N is the number of patches and

$$\boldsymbol{Z}_l^{'} = MSA(LayerNorm(\boldsymbol{Z}_{l-1})) + \boldsymbol{Z}_{l-1}, \tag{6}$$
$$\boldsymbol{Z}_l = MLP(LayerNorm(\boldsymbol{Z}_l^{'})) + \boldsymbol{Z}_l^{'}, \tag{7}$$

for $1 \leq l \leq L$. In Eq. (5), $\boldsymbol{z}_{L0}$ is a vector holding all the useful information and is finally passed through a linear layer to produce the classification vector.

Model ID	MLP Size	Embed Dimension
I	384	192
II	96	96
III	48	48

IV. EXPERIMENTS AND RESULTS

We evaluate our proposed framework on all the 65 various hand gestures of the dataset and considered 3 distinct models, in which MLP size and the Embedding dimension are different. For each model a window size of 64 samples (31.25 ms) is tested to assess the impact of increasing the window size on performance of ViTs. The patch size, models' depth and the number of heads in all of the models are set to $(4, 4)$, 1 , 12 respectively. Adam optimization method is deployed with (β_1 , β_2) equal to $(0.9, 0.999)$, with the learning rate of 0.0001 and the weight decay of 0.001. We fix the batch size and the number of epochs to 128 and 30 respectively and use the Cross-entropy loss function for calculating the models' performance. The Model IDs and their corresponding parameters are presented in Table I.

Table II demonstrates the average accuracy over 19 subjects for each repetition, the average accuracy after performing 5-fold cross validation (i.e., Acc. F1 to Acc. F5) and the corresponding Standard Deviation (STD) for each model. It also shows how many parameters are trained for each specific model. As can be seen, the highest accuracy, in general, pertains to the 3rd repetition and the lowest to the 1st one. The average accuracy rises by 0.38% from model I to model II although the number of parameters in the former is roughly 4 times as large as that in the latter. This indicates that to obtain decent accuracy in ViTs, there is no need to increase the number of parameters and hence the complexity and the training time when this accuracy is achieved with almost $78,000$ parameters. The STD also decreases by a minimal amount when increasing the number of parameters, leading to a reasonable trade-off between the number of parameters on the one hand and the acquired accuracy and STD on the other hand. Fig. 2 visualizes the results from Table II.

For comparison purposes, the proposed ViT-based framework is evaluated against the LDA approach, which is among the most popular conventional ML algorithms that has been widely used for sEMG gesture recognition. We should mention that as the utilized dataset has been released very recently, there are only couple of other works [25], [27] developed based on this dataset. Reference [25] focused on the same task as our work but used traditional ML methods such as LDA. The test-train split, however, is not mentioned in Reference [25] rendering direct comparison inapplicable. Reference [27], on the other hand, only focused on the dynamic and transient phase of gesture movements when the signals are not stabilized or plateaued, which is a different task as this work. Consequently, to have a fair comparison, we have implemented an LDA method similar
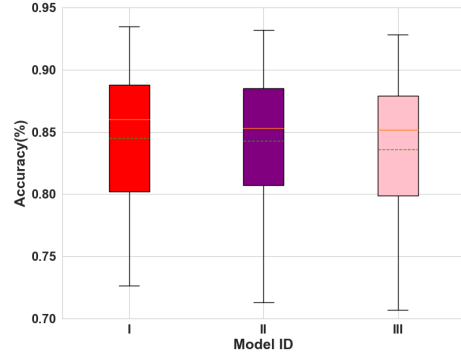

Fig. 2. Accuracy boxplots of 3 different models of the ViT-HGR framework. Each boxplot represents the Interquartile Range for 19 subjects. The accuracy for each subject is the average accuracy after performing 5-fold cross validation.

to that of [25] based on the same setting as our proposed ViT framework. Five key features for classification of sEMG signals including Mean Absolute Value (MAV), the number of Zero Crossings (ZC), Waveform Length (WL), Root Mean Square (RMS) and Slope Sign Change (SSC) along with four auto regressive coefficients of each cropped window are fed to the LDA algorithm [25], [26]. In Table III, the above-mentioned results obtained from the LDA model are presented. Evidently, the average accuracy in the ViT is around 5% bigger and STD is around 8% smaller than that in the LDA, which highlights the great power of ViTs in solving HD-sEMG hand movement classification problems. Furthermore, the signal processing + training time for both ViT-HGRand LDA and for each repetition of one subject is measured. This time is 168 seconds for the ViT-HGR model I and 367 seconds for the LDA, which means for achieving the total average accuracy over all the 19 subjects, we require 4.4 hours while this will be 9.6 hours for the LDA.

V. CONCLUSIONS

In this paper, for the first time (to the best of our knowledge), we introduced a vision transformer-based framework, referred to as the ViT-HGR, for application on HD-sEMG signals for the task of hand gesture classification. To implement the ViT-HGR framework, we capitalize on the recent breakthrough of Transformers in different ML domains and their potentials for employing more input parallelization, therefore, reducing complexity of the underlying model. As direct application of ViT to HD-sEMG is not straightforward, a particular signal processing step is developed to convert the HD-sEMG signals to a specific format that is compatible with ViTs. The proposed ViT-HGR framework can overcome the training time problems associated with recurrent networks and can accurately classify a large number of hand gestures from scratch without any need for data augmentation and/or transfer learning. By comparing the test accuracy associated with three unique variants of the ViT-HGR network, we showed that it could reach average accuracy of 84.62% (for 65 gestures over 19 participants) with no more than $78,000$ parameters. The average accuracy for the LDA is 8% lower and its signal processing + training time is more than

TABLE II

A COMPARISON OF THE AVERAGE ACCURACY FOR EACH FOLD OVER 19 PARTICIPANTS, THE OVERALL ACCURACY AND THE OVERALL STD FOR EACH VIT MODEL.

Model ID	Acc. F1 (%)	Acc. F2	Acc. F3	Acc. F4	Acc. F5	Avg. Acc.	STD (%)	# Parameters
I	75.92	87.79	88.47	87.71	83.22	**84.62**	3.07	340,866
II	75.21	87.34	88	87.88	82.78	**84.24**	3.14	78,210
III	73.92	87.09	87.56	87.09	81.65	**83.46**	3.17	25,314

TABLE III

A COMPARISON OF THE AVERAGE ACCURACY FOR EACH REPETITION OVER 19 PARTICIPANTS, THE OVERALL ACCURACY AND THE OVERALL STD FOR THE LDA MODEL.

Acc. F1 (%)	Acc. F2	Acc. F3	Acc. F4	Acc. F5	Avg. Acc.	STD (%)
82.58	69.65	84.21	82.74	75.27	**78.89**	11.15

twice that of the ViT-HGR. This illustrates potentials of the proposed ViT-HGR framework to act as a feasible substitute for LDAs in hand gesture recognition tasks. This is because the ViTs are more straightforward to be implemented on HD-sEMG data sets with no need for any additional feature extraction calculations and it also takes far less training time, which is a significantly critical issue when working with large data sets. Our primary focus in this paper was on 64-sample portions of the flexor signals because our purpose was to assess the proposed network's performance on small patterns of the HD-sEMG dataset. In the future, we aim to extend the number of channels to evaluate its impact on the framework's efficacy.

REFERENCES

[1] D. Farina, A. Mohammadi, T. Adali, N. V. Thakor, and K. N. Plataniotis, "Signal Processing for Neurorehabilitation and Assistive Technologies" *IEEE Signal Processing Magazine,* vol. 38, no. 4, pp. 5-7, 2021.

[2] S. Tam, M. Boukadoum, A. Campeau-Lecours, and B. Gosselin, "Intuitive Real-time Control strategy for High-density Myoelectric Hand Prosthesis Using Deep and Transfer Learning." *Scientific Reports*, vol. 11, no. 1, pp. 1-14, 2021.

[3] J. Vogel, and A. Hagengruber, "An sEMG-based Interface to Give People with Severe Muscular Atrophy Control over Assistive Devices." *IEEE Int. Conf. of the Engineering in Medicine and Biology Society (EMBC)*, pp. 2136-2141, 2018.

[4] M. Zia ur Rehman, et al., "Stacked Sparse Autoencoders for EMG-based Classification of Hand Motions: A Comparative Multi Day Analyses between Surface and Intramuscular EMG," *Appl. Sci.*, vol. 8, p. 1126, 2018.

[5] E. Rahimian, S. Zabihi, A. Asif, D. Farina, S.F. Atashzar, A. Mohammadi, "FS-HGR: Few-shot Learning for Hand Gesture Recognition via ElectroMyography," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2021.

[6] P. Koch, M. Dreier, M. Maass, M. Böhme, H. Phan, A. and Mertins, "A Recurrent Neural Network for Hand Gesture Recognition Based on Accelerometer Data," *IEEE Int. Conf. of the Engineering in Medicine and Biology Society (EMBC)*, pp. 5088-5091, 2019.

[7] D. Bai, S. Chen, and J. Yang, "Upper Arm Motion High-density sEMG Recognition Optimization Based on Spatial and Time-frequency Domain Features," *Journal of Healthcare Engineering*, 2019.

[8] I. Ketykó, F. Kovács, and K. Z. Varga, "Domain Adaptation for sEMG-based Gesture Recognition with Recurrent Neural Networks," *IEEE Int. Joint Conf. Neural Networks (IJCNN)*, pp. 1-7, 2019.

[10] M. Houston, A. Wu, Y. and Zhang, "Optimizing Input for Gesture Recognition using Convolutional Networks on HD-sEMG Instantaneous Images," *IEEE Int. Conf. of the Engineering in Medicine and Biology Society (EMBC)*, pp. 6539-6542, 2021.

[9] U. Kuruganti, A. Pradhan, and J. Toner, "High-Density Electromyography Provides Improved Understanding of Muscle Function for Those with Amputation," *Frontiers in Medical Technology*, vol. 41, 2021.

[11] J. Chen, S. Bi, G. Zhang, and G. Cao, "High-density Surface EMG-based Gesture Recognition Using a 3D Convolutional Neural Network." *Sensors*, vol. 20, no. 4, p. 1201, 2020.

[12] X. Jiang, X. Liu, J. Fan, et al., "Open Access Dataset, Toolbox and Benchmark Processing Results of High-Density Surface Electromyogram Recordings." *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1035-1046, 2021.

[13] S. Shen, K. Gu, X. R. Chen, M. Yang, and R. C. Wang, "Movements Classification of Multi-channel sEMG Based on CNN and Stacking Ensemble Learning," *IEEE Access*, vol. 7, pp. 137489-137500, 2019.

[14] A. Gautam, M. Panwar, A. Wankhede, S. P. Arjunan, G. R. Naik, A. Acharyya, and D. K. Kumar, "Locomo-net: A Low-complex Deep Learning Framework for sEMG-based Hand Movement Recognition for Prosthetic control," *IEEE J. Transl. Eng. Health Med.*, vol. 8, pp. 1-12, 2020.

[15] T. Triwiyanto, I. P. A. Pawana, and M. H Purnomo, "An Improved Performance of Deep Learning Based on Convolution Neural Network to Classify The Hand Motion by Evaluating Hyper Parameter," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 7, pp. 1678-1688, 2020

[16] W. Wei, *et al.*, "A Multi-stream Convolutional Neural Network for sEMG-based Gesture Recognition in Muscle-computer Interface," *Pattern Recognition Letters*, 2017.

[17] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A Novel Attention-based Hybrid CNN-RNN Architecture for sEMG-based Gesture Recognition," *PloS One*, vol. 13, no. 10, 2018.

[18] E. Rahimian, S. Zabihi, S. F. Atashzar, A. Asif, A. Mohammadi, "Surface EMG-Based Hand Gesture Recognition via Hybrid and Dilated Deep Neural Network Architectures for Neurorobotic Prostheses," *Journal of Medical Robotics Research*, pp. 1-12, 2020.

[19] P. Tsinganos, B. Cornelis, J. Cornelis, B. Jansen, and A. Skodras, "A Hilbert Curve Based Representation of sEMG Signals for Gesture Recognition," *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 201-206, 2019.

[20] A. Vaswani, N. Shazeer, J. Uszkoreit, L. Jones, A. Gomez N., L. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, pp. 5998-60, 2017.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] N. Malešević, A. Olsson, P. Sager, et al., "A Database of High-density surface Electromyogram Signals Comprising 65 Isometric Hand Gestures," *Scientific Data,* vol. 8, no. 1, pp. 1-10, 2021.

[23] E. Rahimian, S. Zabihi, F. Atashzar, A. Asif, A. Mohammadi, "XceptionTime: Independent Time-Window XceptionTime Architecture for Hand Gesture Classification," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[24] B. Hudgins, P. Parker, and R.N. Scott, "A New Strategy for Multifunction Myoelectric Control," *IEEE Trans. Biomed. Eng.* vol. 40, no. 1, pp. 82-94, 1993.

[25] R. N. Khushaba, and K. Nazarpour, "Decoding HD-EMG Signals for Myoelectric Control - How Small Can the Analysis Window Size be?," *IEEE Robotics & Automation Letters,* vol. 6, no. 4, 2021.

[26] M. Atzori, and H. Müller, "PaWFE: Fast Signal Feature Extraction Using Parallel Time Windows." *Frontiers in Neurorobotics,* vol. 13, p. 74, 2019.

[27] T. Sun, Q. Hu, P. Gulati, and S.F. Atashzar, "Temporal Dilation of Deep LSTM for Agile Decoding of sEMG: Application in Prediction of Upper-limb Motor Intention in NeuroRobotics.," *IEEE Robotics and Automation Letters*, 2021.