Hand Gesture Recognition via Transient sEMG Using Transfer Learning of Dilated Efficient CapsNet: Towards Generalization for Neurorobotics

Eion Tyacke, Shreyas P. J. Reddy, Natalie Feng, Rama Edlabadkar, Shucong Zhou, Jay Patel, Qin Hu, *Graduate Student Member, IEEE*, and S. Farokh Atashzar, *Senior Member, IEEE*

***Abstract*—There has been an accelerated surge in utilizing the deep neural network to decode central and peripheral activations of the human nervous system to boost the spatiotemporal resolution of neural interfaces used in human-centered robotic systems, such as prosthetics, and exoskeletons. Deep learning methods are proven to achieve high accuracy but are also challenged by their assumption of having access to massive training samples. *Objective:* In this letter, we propose Dilated Efficient CapsNet to improve the predictive performance when the available individual data is minimal and not enough to train an individualized network for controlling a personalized robotic system. *Method:* We proposed the concept of transfer learning for a new design of the dilated efficient capsular neural network to relax the need of having access to massive individual data and utilize the field knowledge which can be learned from a group of participants. In addition, instead of using complete sEMG signals, we only use the transient phase, reducing the volume of training samples to 20% of the original and maximizing the agility. *Results:* In experiments, we validate the performance with various amounts of injected personalized training data (from 25% to 100% of transient phase). The results support the use of the proposed transfer learning approach based on the dilated capsular neural network when the knowledge domain learned on a small number of subjects can be utilized to minimize the need for new data from new subjects. The model focuses only on the transient phase which is a challenging neural interfacing problem.**

***Index Terms*—Human-centered robotics, neurorobotics, surface electromyography, transfer learning.**

Manuscript received 25 February 2022; accepted 24 June 2022. Date of publication 15 July 2022; date of current version 21 July 2022. This letter was recommended for publication by Associate Editor G. Salvietti and Editor J.-H. Ryu upon evaluation of the reviewers' comments. This work supported in part by the U.S. National Science Foundation under Grants 2037878, 2031594, and 2121391, and in part by NYUAD Center for Artificial Intelligence and Robotics (CAIR) under Award CG010. *(Corresponding author: S. Farokh Atashzar.)*

Eion Tyacke, Natalie Feng, Shucong Zhou, Jay Patel, and Qin Hu are with the New York University, New York, NY 10012 USA (e-mail: et1799@nyu.edu; hf2309@nyu.edu; sz2417@nyu.edu; jp5207@nyu.edu; qh503@nyu.edu).

Shreyas P. J. Reddy is with the International Institute of Information Technology, Bhubaneshwar, Bhubaneswar, Odisha 751029, India (e-mail: b419056@iiit-bh.ac.in).

Rama Edlabadkar is with the Indian Institute of Technology Indore, Simrol, Madhya Pradesh 453552, India (e-mail: mems190005033@iiti.ac.in).

S. Farokh Atashzar is with the Department of Electrical and Computer Engineering, Department of Mechanical and Aerospace Engineering, NYU WIRELESS, NYU Center for Urban Science and Progress (CUSP), New York University, New York, NY 10012 USA (e-mail: f.atashzar@nyu.edu).

Digital Object Identifier 10.1109/LRA.2022.3191238

I. Introduction

In THE U.S., nearly 2 million people are living with limb loss [1], and by estimation, the total number of amputees will approximately exceed 4 million within 30 years [2]. Among all amputations, upper limb loss has a less frequent occurrence but is reported to have a higher rejection rate on commercial prostheses [3], [4]. For the hardware, the current design of the commercial prostheses lacks functionality for Activities of Daily Living (ADLs) and comfortability in wearing. For the software, gesture detection from prostheses deviates from amputees' perception [5]. Periodic, long-duration data collection at designated clinics for maintaining the performance of prostheses is also pushing amputees back to conventional cable-driven, passive prostheses. Thus there is an accelerated surge of research on designing new neural interfaces that can provide more accurate detection of amputees' intended movements and a more agile response to amputees' muscle-activity signal.

In the literature, hand gesture classification through surface electromyography (sEMG) signals [6]–[12] has been extensively utilized with the ultimate goal of use in prosthetic control in a non-invasive manner. However, traditional machine-learning algorithms [7]–[10], which are based on feature engineering, can achieve relatively low accuracy when classifying a large number of gestures. Recently, deep learning (DL) models [11]–[16] were introduced to improve the performance of sEMG-based gesture recognition tasks. Among neural network systems, Convolutional Neural Networks (CNNs) [11], [12] help eliminate the need for manual feature extraction and improve the model performance. Recurrent Neural Networks (RNN) [13], [14] have also been considered due to their ability to capture temporal dynamics. The hybrid models that incorporate CNN and RNN are also applied to leverage the advantages of both techniques [15]. However, deep learning methods can only achieve high performance when trained on a substantial amount of data. Also, it should be noted that by nature, the neural drive to muscles is time-dependent and stochastic, and the neural control strategies between different users vary. In addition, the amputation conditions also cause variability [17]. Consequently, the models trained on specific subjects will not perform well for another user. As a result, the existing models are mainly trained in a user-specific manner, which means that there is a need for large data collection and retraining for new

users. Even for one subject, there will be the need for frequent recalibration (data collection and retraining). Frequent retraining and recalibration [18] is theoretically one way to compensate for the issues; however, the inconvenience it brings also prompts the high rejection rate in the upper limb neurorobotic systems. Due to these reasons, the practicality of models which would require a high volume of sEMG data is challenged, and there is a strong desire to have models that can reduce the need for a high volume of data, even if they support lower degrees of freedom.

To augment the performance, some researchers have leveraged the high spatio-temporal resolution of high-density surface electromyography (HD-sEMG) collected from dense electrode grids [19] to increase the amount of data collected and the information rate in one data collection.

In this letter, we focus on using sparsely-located bipolar sEMG (not high-density) and propose a novel transfer learning algorithm. In the literature, transfer learning is known as a method to deal with data deficiency while ensuring classification accuracy [20], [21]. This letter proposes the use of transfer learning to allow deep learning models pre-train on the sEMG data from multiple users and capture the domain knowledge; then, this pre-trained model will be partially frozen, leaving only a small number of trainable parameters when calibrating on an unseen user. In this way, transfer learning vastly reduces the training samples needed and improves the model performance by transferring the common domain knowledge gained in advance to the calibration tasks for new users. There exists some work on sEMG-based transfer learning using convolutional neural networks [20], [21]. Although it reduces the need for retraining and recalibration, the use of a classic convolutional neural network still needs a large amount of data to achieve high performance and generalizability. Thus, our goal is to find a method that is able to achieve high performance while giving a limited amount of data for training.

In this letter, we conduct transfer learning on a specifically-designed Dilated Efficient Capsule Neural Network, proposed in this letter, as the model architecture. Instead of conventional CNNs that only search for the appearance of learned features from inputs, we use a capsular neural network (CapsNet) to save learned features as vectors with both scalar values and orientation, detecting hand gestures regardless of the "location and orientation" of the associated neurophysiological features in the sEMG data space. In this work, we propose "Dilated Efficient CapsNet" (different from classic CapsNet). The proposed Dilated Efficient CapsNet (a) removes the decoding block that reconstructs the inputs, making the architecture very more compact (reducing more than 89.57% trainable parameters, which also helps with avoiding model overfitting); and (b) uses dilation in each convolutional layer of the convolutional block to enlarge the receptive field in gesture detection. In order to evaluate the robustness and superiority of the proposed model, a systematically-designed comparative study is conducted with multiple conventional deep neural networks.

It should be noted that this letter focuses on utilizing only the transient phase of the sEMG signals from each repetition to reduce control delay. In this regard, it should be mentioned that each repetition of gesture performance in commonly used sEMG databases can be divided into the transient phase, the steady-state, and the descending phase. Transient data corresponds to the bursts of myoelectric activity triggered by sudden motor unit recruitments and indicates the movement intention. In most of the existing literature [6]–[18], [20]–[23], a complete temporal profile (including transient phase, steady-state, and descending phase) or only temporally-dissected steady-state signals (associated with the myoelectric signal during stable muscle contractions) have been utilized for training and validating machine learning models processing sEMG. Thus, transient phase signals are often ignored due to their unstable appearance, even though it plays a critical role in the responsiveness of the neural interfaces. Nevertheless, transient phase signals are observed to possess a unique structure [24], suggesting orderly recruitment of motor units and the potential of including descriptive information of intended movements [25]. In other words, although decoding transient sEMG can be more technically challenging due to the dynamicity of this temporal phase; however, it can significantly improve the quality of the interface. As a result, in this letter, we focus on transient phase sEMG. It should also be noted that utilizing the transient signals reduces the available training data (to at least 20% of the original amount), but it can be used to "predict" the hand gesture before the stable temporal phase. In this letter, gesture prediction is defined as gesture recognition during the dynamic period of gesture performance before the most stable temporal phase of muscle contractions, reducing the control delay to make the interface as seamless as possible. As the major achievement, it can be mentioned that the proposed architecture in this letter requires five times less amount of training data needed for a new subject when compared with state-of-the-art literature [11], [12], [19], [23] while achieving high-performance of about 80% on transient-phase signals (see Table I). The main contributions of this work are as follows:

TABLE I
COMPARISON BETWEEN PROPOSED METHOD AND LITERATURE IN MODEL ACCURACY AND PERCENTAGE OF TRAINING DATA

Paper	# Reps	# Training Reps	Training Data (%)	Test Acc (%)
[11]	6	4	66.67 (4/6*100%)	83.8
[12]	6	4	66.67 (4/6*100%)	78.9
[23]	6	4	66.67 (4/6*100%)	82
[19]	6	4	66.67 (4/6*100%)	82.2
Our Paper	6	4	**13.33 (4/6*20%)**	78.3

Note: All papers are based on Ninapro DB2 Exercise B.

Contribution 1: This letter proposes Dilated Capsule Convolutional Networks in sEMG-based hand gesture prediction for the first time, evaluating the performance of the model for various injections of the new data. By saving vectors instead of scalars, the CapsNet structure can accurately detect unique underlying physiological information associated with each gesture without substantial training data.

Contribution 2: Using different percentages of transient phase signals as training data, we reduce the amount of the training sample to less than 20% (as low as %5) of the original need while significantly enhancing the agility and temporal resolution.

Fig. 1. Total 17 gestures in DB2, Exercise Set B.

Contribution 3: Applying transfer learning on Dilated Efficient CapsNet, we can achieve about 80% accuracy based on transient-phase signals (only 20% of each repetition) when predicting 17 gestures. Moreover, we train on even fewer data to evaluate the performance robustness of the proposed method. As a result, the proposed method can achieve about 70% accuracy when given only 5% signals from each repetition, significantly enhancing the agility and temporal resolution.

II. Data

A. Data Acquisition

Ninapro [26] is an open-source project that aims to help EMG prosthetics research through the publicly available sEMG dataset. In this work, we leverage its second sub-database (DB2), and use the section of "Exercise set B". This dataset includes the sEMG signals [27] of 40 subjects (all intact; 28 males, 12 females; 34 right-handed, 6 left-handed; age 29.9 $\pm$ 3.9 years) performing 17 hand gestures shown in Fig. 1 (8 isometric and isotonic hand configurations, and 9 basic movements of the wrist). To collect the data, each subject was asked to maintain the hand gesture for 5 seconds, followed by resting for 3 seconds, and the experiment was repeated 6 times for each gesture. In addition, the acquisition setup included 12 Delsys Trigno electrodes (8 electrodes were wrapped around the radio-humeral joints, two around the biceps and triceps, and two around the flexor and extensor digitorum superficialis). They were designed to record hand kinematics, dynamics, and the corresponding muscular activity, and the sensors were linked to a laptop responsible for signal data acquisition sampled at 2k Hz. Using such a benchmarked database allows us to ignore various factors, including the experimental conditions which could otherwise affect the recorded results.

B. Data Pre-Processing

The three phases of repetition of voluntary muscle contraction form a trapezoidal profile and are visualized using sEMG accelerometer data. The upper base of the trapezoid is considered

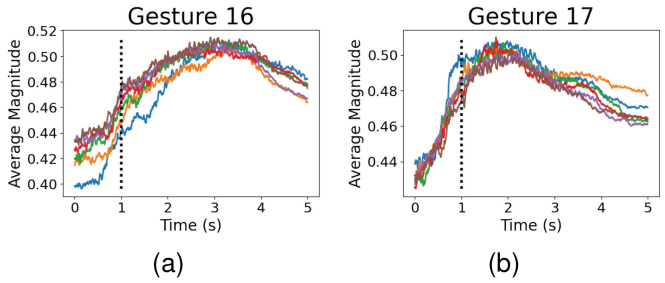

Fig. 2. Illustration of transient phase signals: average root mean square of accelerometer data across all subjects for (a) gesture 16, and (b) gesture 17. The dotted line indicates that the transient phase ends at 1 s. Repetitions are denoted by different colors.

the steady-state of contraction, and the first slant of the trapezoid roughly outlines the transient phase. In this letter, we decide the length of the transient phase by the root mean square (RMS) of the accelerometer signals averaging across all subjects. As shown in Fig. 2, it is observed that the average RMS of the accelerometer data becomes steeper in the first 20% (1 second on average) of the gesture repetition. This part can be considered the transient phase, and thus, we extract the first 20% of each repetition length to obtain the transient, discarding the remaining part of the repetition.

A minimal preprocessing pipeline that includes normalization and rectification is then processed on the data. It normalizes the signals using z-score normalization with means and standard deviations from the training data and rectifies the normalized signals by taking the absolute values. Normalized and rectified signals are windowed with 300 ms, and labels are assigned to each window. The signal data after windowing will be in the shape of 600 * 12. The 600 here represents the number of timesteps (300 ms * 2 kHZ = 600), and 12 represents the number of sensors or channels. Based on the literature, 300 ms is the largest window size required for real-time control. We use a window of size 300 ms [24] with an overlap of 10 ms to generate the training and testing sets. Data in the DB2 dataset is already Hampel filtered to remove 50 Hz powerline interference [28]. We do not apply any extra lowpass filtering techniques to the signals since it reduces the quality of the data. For the train-test split, we leverage repetitions 1, 3, 4 and 6 for training the deep learning model and repetitions 2,5 for validating the trained model [28]. Therefore, our proposed method only requires a total duration of 68 seconds (4 repetitions * 1 second per repetition * 17 gestures) for each subject for training. The transient phase of the signals being a small part of the data sample drastically reduces the time required to train the model and results in a lower calibration time.

III. Model Architecture

We propose the concept of Dilated Efficient Capsule Network in this letter for sEMG signals classification and apply transfer learning to it to improve gesture-prediction accuracy and minimize the needed data for training.

In the literature, Convolutional Neural networks (CNN) have helped achieve remarkable results in problems including image

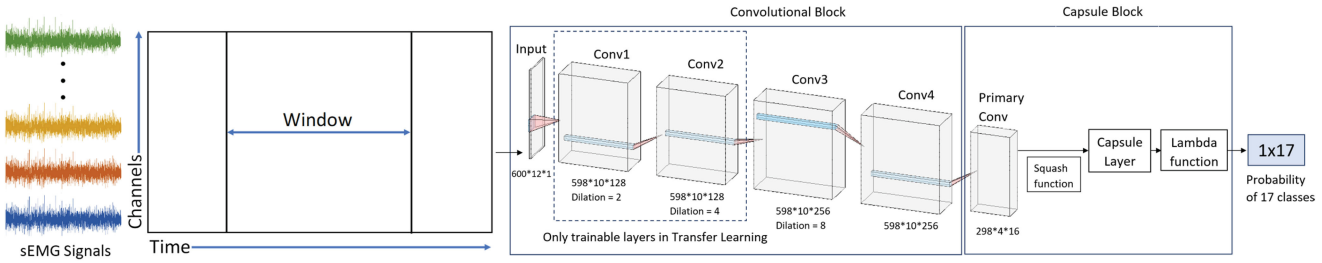

Fig. 3. The Illustration of the model architecture of Dilated Eff-Caps. 1) We process the sEMG signals by windowing and make it the input of Dilated Eff-Caps. 2) The Dilated Eff-Caps consists of two blocks. The Convolutional block has four Conv layers, each having a 2D Convolutional Layer with a ReLU activation function at the end of each layer. The Capsule block has a Primary Conv followed by a squash function, a Capsule layer to perform routing algorithms, and a Lambda Function to generate the output. 3) The output of this model architecture is the probability of each signal classified as one of the 17 gestures. 4) The dashed line denotes the only two trainable Conv layers of the Top5 Eff-Caps model when calibrating individually on the remaining 35 subjects.

classification [29]–[31] to object detection [32]–[34]. However, translation invariance achieved by the CNN comes at the expense of losing some information about the object's location. This issue has been addressed generically in the literature using capsular neural network architectures. Capsules are vector representations of features. Moreover, each capsule involved in the network dynamically describes how the entity is instantiated. Capsule networks also leverage the concept of routing by agreement, where the predictions of low-level capsules are routed to their best match parent. This helps assess the reciprocal agreement between groups of neurons to capture covariance and leads to a compact model with fewer parameters and better capability to generalize on new data.

The model architecture of the proposed Efficient CapsNet has two basic building blocks as described in Fig. 3. The Convolutional block involves four 2D convolutional layers, having 128, 128, 256, and 256 filters, respectively. Each 2D convolutional layer is followed by a Dropout Layer and a Batch Normalization Layer. We use the Dropout regularization technique to avoid the possible case of over-fitting. The dropout rate in our model is set to 0.5. We have the Convolutional block followed by a Capsule block. The Capsule block involves a "Primary Convolutional" layer that has a 2D convolutional layer followed by a squash activation function to normalize the output vector rather than the scalar elements themselves. The squash activation function is defined as (1), where $\mathbf{v}_j$ is the vector output of capsule j, and $\mathbf{s}_j$ is its total input.

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2}\frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \tag{1}$$

It is necessary to ensure that the length of the vector lies in the range from 0 to 1 because it represents the probability of information routing from the current layer to the next capsule layer. Hence a squash activation is used, which drives the length of the large vector towards 1 and the small vector towards 0. The Capsule layer in the Capsule block expands the output of the neuron from a scalar to a vector. We use an iterative dynamic routing algorithm that groups capsules to form a parent capsule to compute the capsule's output. The number of routing iterations is 3. We define the "Margin" loss which is computed for each class in the classification problem, given by (2).

$$L_k = \underbrace{T_k \max\left(0, m^+ - \|\mathbf{v}_k\|\right)^2}_{\text{class present}} + \underbrace{\lambda\left(1-T_k\right)\max\left(0, \|\mathbf{v}_k\| - m^-\right)^2}_{\text{class not present}} \tag{2}$$

T_k is assumed to be 1 if the gesture k is present or 0 otherwise. The parameters m+ and m− are tuned so that the length of the vector does not become extreme. The Lambda function computes the length of each output vector from the Capsule layer. These lengths represent the probability of each of the 17 gestures.

The proposed Dilated Efficient CapsNet is an improvement of the currently existing CapsNet model in literature [35]. The original CapsNet had 16.3M trainable parameters in comparison to the proposed Dilated Efficient CapsNet, which has 1.7M trainable parameters. The primary difference between the two models is their architecture. The original CapsNet has a single convolutional layer in the convolutional block in comparison to the Efficient CapsNet, which has four convolutional layers, each with a large number of filters. To avoid the complexity of the original CapsNet model, the Efficient implementation removes the reconstruction/decoding block of the CapsNet. This reduces the number of trainable parameters and training time significantly. Removing the decoder part does not affect the model's performance because we currently use minimal pre-processed windowed data as the input to the model. To improve the performance of the model, we make use of dilation to expand the area of reach without pooling. We use a dilation rate of (2, 2) in the 2nd convolutional layer followed by (4, 4) in the 3rd layer and (8, 8) in the last layer of the convolutional block. The total number of trainable parameters for the Dilated Efficient CapsNet becomes 3.7M, however, the additional padding incorporated with dilation allows the center of the kernel to pass over the edge channels, providing additional useful information. The training process in a real-life situation can be quite time-consuming.

To solve the data hunger problem of DL-based gesture recognition for decoding multichannel sEMG, we leverage the concept of Transfer Learning and apply it in our proposed Dilated

Efficient CapsNet (Dilated Eff-Caps) model. The Transfer learning model uses the pre-knowledge from other users and has a reduced number of trainable parameters when calibrating on an unseen user. The quality of data used to train the pre-trained model for transfer learning affects the representativeness and generalizability of the pre-trained model when calibrating on new users. Hence, a poor subset would not accurately represent the inter-subject variance and would not be very useful. In our study, the proposed Dilated Eff-Caps model is initially trained on each of the 40 subjects in the dataset, and the corresponding performances were recorded to derive the top-5 performing subjects from the dataset for transfer learning. Then, we pre-train the Dilated Eff-Caps on the data from the top-5 performing subjects together, leaving only the first two convolutional layers (see the dashed area in Fig. 3) of the pre-trained Dilated Eff-Caps trainable. With the common knowledge of the 5 best-performing subjects, the calibrated model continuously learns from and tailors to a new user when there is fresh information coming in. In this way, we reduce the negative impact of less training data and improve training efficiency.

IV. Experiments and Results

A set of experiments are conducted to validate our proposed model architecture and the segmentation scheme of the transient phase. We quantify the advantages of transfer learning (see Appendix A) and trade-offs presented through the use of the transient signals (see Appendix B) before comparing the performance of the Dilated Eff-Caps and Top5 Eff-Caps with a multi-layer perceptron (MLP), 2D CNN, and RNN-CNN Hybrid [23]. Evaluation of the predictive accuracy is carried out through the use of different transient segmentation schemes with different data volumes.

The Hybrid model architecture, for comparison, was inspired by [23]. The hybrid model has three components: the LSTM block composed of four LSTM layers, each with 128 units; the CNN block with seven convolutional layers with the filters of 32, 64, 64, 128, 128, 256, respectively and the dilation rate on CNN layers two through six of 2, 2, 4, 8, 8, 8; and the classifier block, three fully connected dense layers with units 64, 32, 17. The size of the MLP and 2D CNN network was created to be comparable to the number of trainable parameters in the Hybrid model 1.7M. The MLP architecture used is a simple five-layer network with the following number of units per layer: 256, 128, 64, 32, 17. The 2D CNN was composed of four 2D convolutional layers with filters of 32, 32, 64, and 64. The convolutional block was then followed by a small classifier block of two dense layers with 32 and 17 units.

Regarding the data segmentation scheme, we propose (a) temporal segmentation, and (b) repetition segmentation. As shown in Fig. 4, we split sEMG signals by repetition 1, 3, 4, 6 and repetition 2, 5. Purple repetitions 1, 3, 4, 6 are fed into the models as training samples, and red repetitions 2, 5 are served as the testing data. To validate the proposed model's ability to achieve high predictive accuracy with much less training data, we extract different percentages of transient signals as model input. Temporal segmentation extracts signals via horizontal

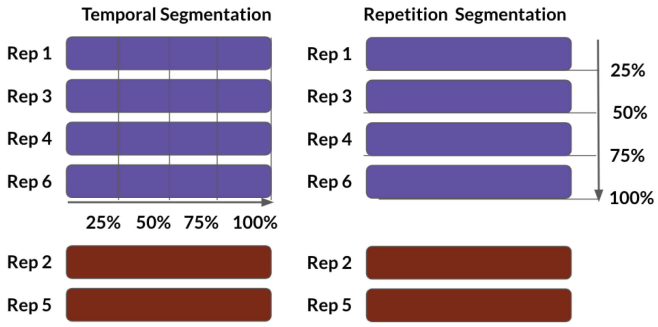

Fig. 4. The illustration of data segmentation scheme: temporal segmentation and repetition segmentation. Repetitions in purple are training samples, and repetitions in red are testing samples. We extract different percentage of the transient phase via the direction of arrow.

TABLE II
Temporal Segmentation (Average Accuracy)

Temporal Percentage	MLP	2D CNN	Hybrid	Dilated Eff-Caps	Top5 Eff-Caps
25%	0.484	0.429	0.502	0.565	0.694
50%	0.567	0.529	0.562	0.701	0.740
75%	0.616	0.603	0.639	0.761	0.777
100%	0.647	0.651	0.713	0.775	0.783

TABLE III
Repetition Segmentation (Average Accuracy)

Repetition Percentage	MLP	2D CNN	Hybrid	Dilated Eff-Caps	Top5 Eff-Caps
25%	0.360	0.334	0.359	0.381	0.367
50%	0.543	0.511	0.513	0.603	0.605
75%	0.601	0.584	0.606	0.680	0.663
100%	0.647	0.651	0.713	0.775	0.783

chronological order. In contrast, repetition segmentation sees each repetition as 25% of the training transient signals and extracts a various number of repetitions based on experimental need via vertical order.

The comparative average model performance of temporal segmentation is included in Table II, and repetition segmentation model comparison is included in Table III. Both tables correspond with a box plot to visualize the accuracy distribution of the 35 tested subjects from all models. A two-sided Mann-Whitney-Wilcoxon test was applied with Bonferroni correction. We set the p-value to be 0.05. Significance is indicated on results using the following significance markers: corrected p-values between 0.05 and 1 are considered to be not significant (ns); corrected p-values between 0.01 and 0.05 are marked by *; corrected p-values between 0.001 and 0.01 are marked by **; corrected p-values between 0.0001 and 0.001 are marked by ***, and corrected p-values smaller than 0.0001 are marked by ****. The statistical significance test and p-value annotations apply to all model comparisons in the letter, including the Appendix section.

Overall, we can notice a higher predictive accuracy of temporal segmentation experiments than repetition segmentation

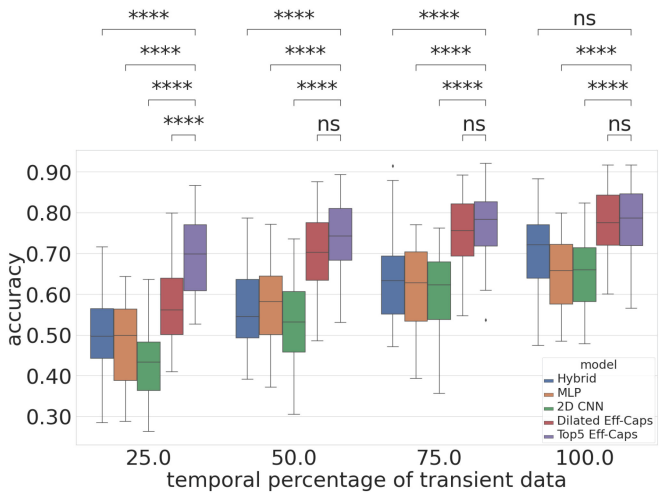

Fig. 5. Temporal segmentation performance comparison with statistical significance tests across selected models.

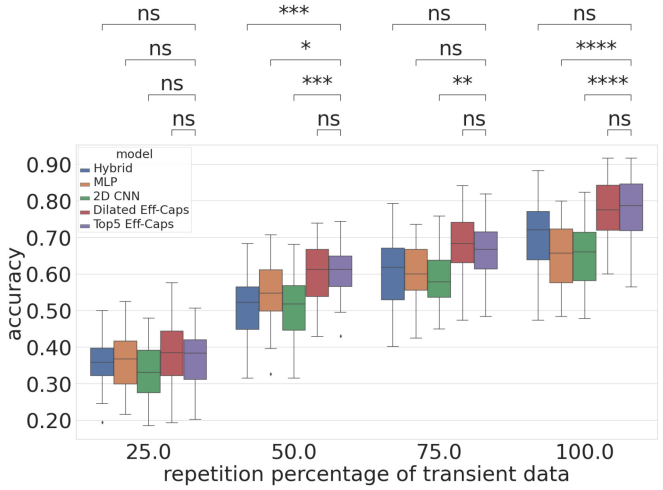

Fig. 6. Repetition segmentation performance comparison with statistical significance tests across selected models.

experiments. All models' predictive accuracy of repetition segmentation tasks is highly correlated to the amount of data fed as training samples, while in temporal segmentation, Dilated Eff-Caps and Top5 Eff-Caps can still achieve high accuracy with much fewer data. It can be observed by comparing Figs. 5 and 6 that our proposed method (Top5 Eff-Caps) can achieve as higher as more than 30% (69.4%–36.7%) accuracy in temporal segmentation than in repetition segmentation when given the same amount of training data (25% of transient data or 5% of complete data). This observation matches our expectation because, the fewer data required from each repetition, the lower the control delay. Based on temporal segmentation, Top5 Eff-Caps can achieve 69.4% predictive accuracy with only 25% of the transient (5% of the complete signals), and achieve 74%, 77.7%, 78.3% with 50%, 75%, and 100% percentage data respectively. The Top5 Eff-Caps net outperforms any other model architecture in terms of the predictive accuracy with any transient percentage. At most, it can achieve 43.4% higher accuracy with 25% of transient data compared to MLP; at least, it can achieve 10% higher accuracy with 75% of transient data compared to the Hybrid model. It can be seen that although the Top5 Eff-Caps is not significantly better than its non-transfer-learning counterpart when using over 50% of the transient signals and is not significantly better than the hybrid model when using 100% of the transient signals, Top5 Eff-Caps outperforms the other models when using only 25% of the transient signals.

Based on repetition segmentation, the effectiveness of transfer learning is not as dominant as in the case of temporal segmentation. The statistical significance test highlights this relative reduction in superiority. The Top5 Eff-Caps model at the 25% and 75% repetition segmentation levels obtains lower accuracies than the non-transfer-learning Dilated Eff-Caps model. However, our proposed models, Dilated Eff-Caps and Top5 Eff-Caps net, outperform all other comparative models, validating the effectiveness of the architecture in this letter.

In order to further evaluate the results of each segmentation method, the variance in the accuracy for all models was checked. The variances of Top5 Eff-Caps model are the lowest given any amount of training data in temporal segmentation, with an average variance of 7.47%. For repetition-based segmentation, the variances of the Top5 Eff-Caps net model is the lowest given all amounts of training data except 25%, with an average variance of 5.67%. These results further support the benefit of transfer learning and the performance of the proposed algorithm.

V. Conclusion

This letter proposes transfer learning for a new Dilated Efficient CapsNet, designed in this letter, to optimize the gesture prediction accuracy with much fewer data provided compared to conventional studies. Instead of using complete sEMG signal repetitions, this work utilizes only the transient phase as the training samples. In this way, we reduce the training data to at least 20% of the original and transform the traditional classification problem into prediction. Utilizing only transient signals can eventually augment the predictive agility of neurorobotics. The inherent capsular characteristics of saving features as vectors of the proposed Dilated Capsule Neural Network helps to capture the correlated information within signals and across repetitions. It ensures the performance of the proposed model architecture with a very small amount of data compared to conventional deep learning methods applied to the sEMG decoding problem. The results showed that the implemented transfer learning approach for the proposed Dilated Efficient CapsNet can achieve an average predictive accuracy of 78.3% across 35 subjects on transient signals (20% of the complete signal repetitions). Even with 25% of the transient (5% of the complete signal repetitions), it can still achieve an accuracy of about 70%.

In this work, we also examine two different data segmentation schemes by comparing the performance of various models under these two circumstances: temporal segmentation and repetition segmentation. By processing these two segmentation schemes, we can check the feasibility of further reducing the training data by 25%, 50%, and 75% of its original. A superiority of temporal segmentation over repetition segmentation is observed, and the

effectiveness of transfer learning under temporal segmentation is larger than that under repetition segmentation. Although we use much fewer data than in previous research, it is noteworthy to mention that the data used in this letter had been collected from able-bodied subjects. The neurophysiology of a healthy population can hardly reflect amputees' bio-electrical conditions, and various amputations would also lead to greater challenges of training only on transient signals of non-intact subjects. This is a limitation of the current study. As a future line of research, we will collect the sEMG data from amputees and validate our model architecture in more practical applications. One of the lines of our future work is to investigate adaptive and iterative model learning that reduces the need for periodic retraining and recalibration. This is not within the scope of the current letter. Another future line of our research is to propose a systematic method (e.g., Markov Decision Process) to distinguish the transient phase and steady-state phase when given a repetition of sEMG signal.

Appendix
Method Verification

In this section, we describe and discuss the additional experiments that help us come to the conclusion of having Top5 Eff-Caps as our best proposed model. These experiments answer questions, including 1) "what are the benefits of applying transfer learning to Dilated Eff-Caps?", 2) "what and how many subjects should be used in a pre-trained model?", and 3) "what is the trade-off between high model performance (when training on steady-state or complete data) and low control delay (when training only on the transient data)?". The detailed experiments and observations can be found in the following subsections.

Appendix A
Performance Comparison with and without Transfer Learning

In order to exemplify the benefit of transfer learning with our proposed model architecture, we have provided the following comparison displaying the per-gesture accuracy across 30 subjects' transient data for the Dilated Eff-Caps model and two transfer learning variants, i.e., the Top5 Eff-Caps and the Top5-Rand5 Eff-Caps. Top5-Rand5 Eff-Caps is very similar to Top5 Eff-Caps, but in addition to using the top 5 subjects' data for pre-training, data from five additional randomly selected subjects were included. The Top5-Rand5 Eff-Caps were considered to see whether adding additional data for pre-training without discretion would result in higher user-specific performance for the remaining subjects.

Fig. 7 indicates that without transfer learning, the model has a lower performance accuracy overall and is less robust, as seen by its larger variance. Gestures 1, 3, 8, and 12 exemplify these poorer qualities of the non-transfer learning model. Additionally, the plot portrays the importance of subject data selection as it pertains to model generalizability when implementing transfer learning. Observing the inclusion of the random five subjects in our Top5-Rand5 Eff-Caps model, there is a subtle decrease in performance, evidenced by a slight reduction in overall accuracy

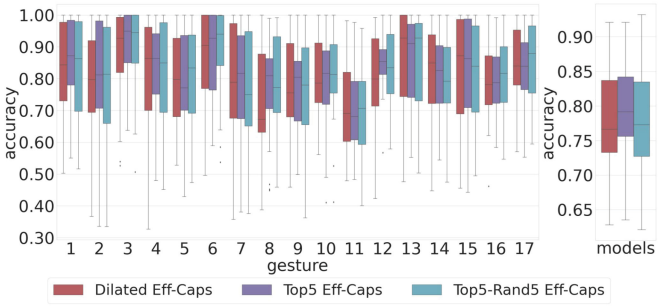

Fig. 7. The box plot on the left side shows the per-gesture accuracy across 30 subjects for each model. The box plot to the right side displays the average accuracy across all gestures for each model.

and an increase in variance. This reduction in performance is reasonable as pretraining was performed with mixed quality subject data, and most of the model layers remained frozen during retraining. The slight reduction in performance is because the larger-scale feature representations found through hierarchical convolutions of multiple layers are not as clear and cleanly defined as those found in the Top5 Eff-Caps.

The model performances from worst to best are non-transfer learning Dilated Eff-Caps with an accuracy of 80.5% and a subject variance of 2.42%; Top5-Rand5 Eff-Caps with an accuracy of 81.3% and a subject variance of 2.28%; Top 5 Eff-Caps with an accuracy of 81.1% and a subject variance of 1.99%. It can be noted that even though the comparing three models similarly perform when given 100% transient data, applying transfer learning on the proposed dilated Eff-Caps enhances the generalizability of a pre-trained model and reduces the discrepancy in user-specific accuracies of the tested 35 subjects.

Appendix B
Performance Control Delay Trade-off Analysis

In subsection B, we analyze the trade-off between low control delay using only the transient-phase signals and the relatively high performance using the steady-state-phase or complete signals. Considering the desired profile of gesture performance is an isosceles trapezoid, in accordance with the observations on the accelerometer signals (see Fig. 2), we consider the first 20% (one second in average) data as transient-phase signals, the next 60% (three seconds in average) data as steady-state-phase signals, and the remaining 20% (one second in average) data as descending-phase signals from each repetition. In this appendix, the model performance of the proposed method (Top5 Eff-Caps) is evaluated in these three phases resulting in 78.1% on transient-phase data, 81.9% on steady-state-phase data, and 81.5% on complete data (see Fig. 8). When implementing the statistical significance test (two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction), we observe no significant difference between the model performance on transient data and the model performance on either steady-state-phase data or complete data. The use of the transient signal provides a substantial improvement in control time while here we show the performance is statically not different. Achieving an accuracy near 80% across

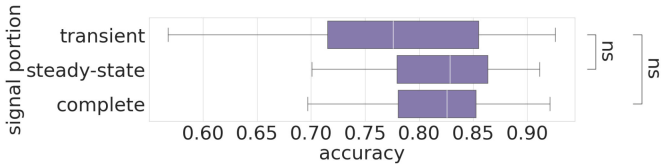

Fig. 8. Top5 Eff-Caps performance when training and testing on different signal portions. The average accuracies across 35 subjects 78.1% based on transient data, 81.9% based on steady-state data, and 81.5% based on complete data.

17 different gestures while using only the transient signal is a significant achievement considering the data limiting constraints. The model performance (∼80%) based on transient-phase data is comparable to the literature, where models were trained based on complete signals from the same data set (Ninapro DB2 Exercise B). Accordingly, it can be said that the proposed method in this letter outperforms the conventional deep learning methods by at least 7% when using minimal data (see Tables II and III).

REFERENCES

[1] K. Ziegler-Graham *et al.*, "Estimating the prevalence of limb loss in the united states: 2005 to 2050," *Arch. Phys. Med. Rehabil.*, vol. 89, no. 3, pp. 422–429, 2008.

[2] L. Resnik *et al.*, "Advanced upper limb prosthetic devices: Implications for upper limb prosthetic rehabilitation," *Arch. Phys. Med. Rehabil.*, vol. 93, no. 4, pp. 710–717, 2012.

[3] M. A. Adwan *et al.*, "Variables affecting amputees' reaction to the artificial limbs in the kingdom of jordan," *J. Amer. Acad. Special Educ. Professionals*, pp. 140–154, 2017.

[4] K. A. Raichle *et al.*, "Prosthesis use in persons with lower-and upper-limb amputation," *J. Rehabil. Res. Develop.*, vol. 45, no. 7, 2008, Art. no. 961.

[5] K. Østlie *et al.*, "Prosthesis rejection in acquired major upper-limb amputees: A population-based survey," *Disabil. and Rehabil.: Assistive Technol.*, vol. 7, no. 4, pp. 294–303, 2012.

[6] F. S. Sayin, S. Ozen, and U. Baspinar, "Hand gesture recognition by using sEMG signals for human machine interaction applications," in *Proc. Signal Process.: Algorithms, Architectures, Arrangements, Appl.*, 2018, pp. 27–30.

[7] J. J. A. M. Junior, M. Freitas, H. Siqueira, A. E. Lazzaretti, S. Stevan, and S. F. Pichorim, "Comparative analysis among feature selection of sEMG signal for hand gesture classification by armband," *IEEE Latin America Trans.*, vol. 18, no. 06, pp. 1135–1143, Jun. 2020.

[8] A. Phinyomark *et al.*, "EMG feature evaluation for improving myoelectric pattern recognition robustness," *Expert Syst. Appl.*, vol. 40, no. 12, pp. 4832–4840, 2013.

[9] M. Tavakoli *et al.*, "Robust hand gesture recognition with a double channel surface EMG wearable armband and SVM classifier," *Biomed. Signal Process. Control*, vol. 46, pp. 121–130, 2018.

[10] A. Gijsberts *et al.*, "Movement error rate for evaluation of machine learning methods for sEMG-based hand movement classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 4, pp. 735–744, Jul. 2014.

[11] W. Wei *et al.*, "A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface," *Pattern Recognit. Lett.*, vol. 119, pp. 131–138, 2019.

[12] Z. Ding *et al.*, "sEMG-based gesture recognition with convolution neural networks," *Sustainability*, vol. 10, no. 6, 2018, Art. no. 1865.

[13] M. Jabbari, R. N. Khushaba, and K. Nazarpour, "EMG-based hand gesture classification with long short-term memory deep recurrent neural networks," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2020, pp. 3302–3305.

[14] M. Hioki and H. Kawasaki, "Estimation of finger joint angles from sEMG using a recurrent neural network with time-delayed input vectors," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, 2009, pp. 289–294.

[15] Y. Hu *et al.*, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS One*, vol. 13, no. 10, 2018, Art. no. e0206049.

[16] K. Shi *et al.*, "MCSNet: Channel synergy-based human-exoskeleton interface with surface electromyogram," *Front. Neurosci.*, vol. 15, 2021, Art. no. 704603.

[17] S. T. P. Raghu, D. T. MacIsaac, and A. D. C. Chan, "Automated biomedical signal quality assessment of electromyograms: Current challenges and future prospects," *IEEE Instrum. Meas. Mag.*, vol. 25, no. 1, pp. 12–19, Feb. 2022.

[18] A. M. Simon *et al.*, "Patient training for functional use of pattern recognition-controlled prostheses," *J. Prosthetics Orthotics: JPO*, vol. 24, no. 2, 2012, Art. no. 56.

[19] T. Sun, Q. Hu, J. Libby, and S. F. Atashzar, "Deep heterogeneous dilation of LSTM for transient-phase gesture prediction through high-density electromyography: Towards application in neurorobotics," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2851–2858, Apr. 2022.

[20] U. Côté-Allard, C. L. Fall, A. Campeau-Lecours, C. Gosselin, F. Laviolette, and B. Gosselin, "Transfer learning for sEMG hand gestures recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2017, pp. 1663–1668.

[21] X. Chen, Y. Li, R. Hu, X. Zhang, and X. Chen, "Hand gesture recognition based on surface electromyography using convolutional neural network with transfer learning method," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 4, pp. 1292–1304, Apr. 2021.

[22] R. Chowdhury, M. B. I. Reaz, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, pp. 12431–12466, 2013.

[23] P. Gulati, Q. Hu, and S. F. Atashzar, "Toward deep generalization of peripheral EMG-based human-robot interfacing: A. hybrid explainable solution for neurorobotic systems," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2650–2657, Apr. 2021.

[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, Jan. 1993.

[25] R. Merletti and P. J. Parker, *Electromyography: Physiology, Engineering, and Non-Invasive Applications*, vol. 11. Hoboken, NJ, USA: Wiley2004.

[26] 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.

[27] M. Atzori *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, no. 1, pp. 1–13, 2014.

[28] K. Englehart, B. Hudgin, and P. A. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 3, pp. 302–311, Mar. 2001.

[29] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 770–778.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[33] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[34] V. Mazzia, A. Khaliq, F. Salvetti, and M. Chiaberge, "Real-time apple detection system using embedded systems with hardware accelerators: An edge AI application," *IEEE Access*, vol. 8, pp. 9102–9114, 2020.

[35] S. Sabour *et al.*, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.