

FS-HGR: Few-shot Learning for Hand Gesture Recognition via ElectroMyography

Elahe Rahimian[†], Student Member, IEEE, Soheil Zabihi[§], Student Member, IEEE, Amir Asif^{††}, Senior Member, IEEE, Dario Farina^{‡‡}, Fellow, IEEE, Seyed Farokh Atashzar[‡], Member, IEEE, and Arash Mohammadi[†], Senior Member, IEEE

[†]Concordia Institute for Information System Engineering (CIISE), Concordia University, Montreal, QC, Canada

[§]Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

^{‡‡} Department of Bioengineering, Imperial College London, London, UK

[‡] Electrical & Computer Engineering and Mechanical & Aerospace Engineering, New York University, USA

^{††}Electrical and Computer Engineering, York University, Toronto, Canada

Abstract—This work is motivated by the recent advances in Deep Neural Networks (DNNs) and their widespread applications in human-machine interfaces. DNNs have been recently used for detecting the intended hand gesture through processing of surface electromyogram (sEMG) signals. The ultimate goal of these approaches is to realize high-performance controllers for prosthetic. However, although DNNs have shown superior accuracy than conventional methods when large amounts of data are available for training, their performance substantially decreases when data are limited. Collecting large datasets for training may be feasible in research laboratories, but it is not a practical approach for real-life applications. Therefore, there is an unmet need for the design of a modern gesture detection technique that relies on minimal training data while providing high accuracy. Here we propose an innovative and novel “Few-Shot Learning” framework based on the formulation of meta-learning, referred to as the FS-HGR, to address this need. Few-shot learning is a variant of domain adaptation with the goal of inferring the required output based on just one or a few training examples. More specifically, the proposed FS-HGR quickly generalizes after seeing very few examples from each class. The proposed approach led to 85.94% classification accuracy on new repetitions with few-shot observation (5-way 5-shot), 81.29% accuracy on new subjects with few-shot observation (5-way 5-shot), and 73.36% accuracy on new gestures with few-shot observation (5-way 5-shot).

Index Terms—Myoelectric Control, Electromyogram (EMG), Meta-Learning, Few-Shot Learning (FSL).

I. INTRODUCTION

The recent advances in Machine Learning (ML) and Deep Neural Networks (DNNs) coupled with innovations in rehabilitation technologies have resulted in a surge of significant interest in the development of advanced myoelectric prosthesis control systems. Hand motion recognition via sEMG signals [1], [2] is considered as a central approach in the literature. Conventional ML techniques, such as Linear Discriminant Analysis (LDA) [3]–[5] and Support Vector Machines (SVMs) [3], [4], [6], have been used for detecting the intended hand gesture through processing of surface EMG (sEMG) signals. Although classical pattern-recognition-based

This Project was partially supported by the Department of National Defence’s Innovation for Defence Excellence and Security (IDEaS) program, Canada.

myoelectric control has been widely studied in academic settings over the last decades, the advanced methodologies have not been used in many commercial examples. This is due to a noticeable gap [1], [7], [8] between real-world challenges and existing methodologies. Among the reasons for this gap are:

- (i) *Training Time*: The first problem is the extended training time required by the end-user to mitigate the differences between the desired and performed movements. Such a training process, which is time consuming, tedious and unpleasant, can take up to several days in practice.
- (ii) *Variability in the characteristics of sEMG Signals*: The second issue is the variability in the nature of the sEMG signals. This variability is caused by (a) Time-dependent and stochastic nature of the neural drive to muscles; (b) Dependency of the neural drive to the dynamic and kinematics of tasks, and; (c) Variability in neural control strategies between different users and the changes caused by amputations. In addition, sEMG recording could vary based on electrode location. Given such variations, therefore, the probability distributions of sEMG signals may be different over time. Consequently, models trained based on some specific observations may not consistently and directly be reused over time. This would require retraining and recalibration, which cannot be done often in real-life applications.

Recently, DNNs have been designed and used by our team [9]–[12] and other research groups [13]–[18], for myocontrol, achieving superior classification performance than conventional approaches. However, DNNs need large training data to achieve high performance. This may be feasible in laboratory conditions but poses constraints in the practical use of prostheses in real-life applications. There is an unmet need for the design of a modern gesture detection technique that relies on minimal training data while achieving high performance.

In this paper we introduce, for the first time, the concept of few-shot training for myoelectric systems. Few-shot learning minimizes the need for recalibration and would allow the user to retrain the ML core of control, by only few basic exercises

instead of extensive recalibration procedures. For this purpose, here we propose an innovative *Few-Shot Learning framework, referred to as the FS-HGR**. The proposed meta-learning FS-HGR architecture takes advantage of domain knowledge and requires a small amount of training data (when compared with traditional counterparts) to decode new gestures of the same or new users. The paper makes the following contributions:

- A class of architectures is introduced for sEMG meta-learning, where the meta-learner, via adaptation, quickly incorporates and refers to the experience based on just few training examples.
- The proposed FS-HGR framework allows a myoelectric controller that has been built based on background data to adapt to the changes in the stochastic characteristics of sEMG signals. The adaptation can be achieved with a small number of new observations making it suitable for clinical implementations and practical applications.
- By proposing the FS-HGR framework, which utilizes a combination of temporal convolutions and attention mechanisms, we provide a novel venue for adopting few-shot learning, to not only reduce the training time, but also to eventually mitigate the significant challenge of variability in the characteristics of sEMG signals.

The paper is organized as follows: Section II provides a brief overview of relevant literature. In Section III, we present the dataset used in development of the proposed FS-HGR framework together with the pre-processing step. The proposed FS-HGR architecture is developed in Section IV. Experimental results and different evaluation scenarios are presented in Section V. Finally, Section VI concludes the paper.

II. RELATED WORKS

A common strategy used for hand gesture recognition in recent works is applying DNN with the focus on improving hand gestures classification performance on “*never-seen-before repetitions*”. Along this line of research, several state-of-the-art works [10]–[12], [14], [16]–[22] mainly used the Niapro database [23]–[25], which is a public dataset providing kinematic and sEMG signals from 52 finger, hand, and wrist movements. The Niapro database is similar to data obtained in real-world conditions, and as such it allows development of advanced DNN-based recognition frameworks.

The common approach in recent studies [10]–[12], [14], [16]–[22], following the recommendations provided by the Niapro database, is to train DNN-based models on a training set consisting of approximately 2/3 of the gesture trials of each subject. The evaluation is then performed on the remaining trials constituting the test set. Although existing DNN techniques achieve promising performance on never-seen-before repetitions, they fail to function properly if the repetition is not extensively explored [26]–[28]. Thus, for a new user or a new gesture, a significant amount of training should be conducted and the whole learning process should be redone, assuming a small variation between the new class and the previous classes.

*The source code of the proposed FS-HGR framework is available at <https://ellarahimian.github.io/FS-HGR/>

If the aforementioned change is more than minimal, there may be the need to recalibrate the whole process for all classes. In addition, existing DNN-based methodologies require large training datasets and perform poorly on tasks with only a few examples being available for training purposes.

In Reference [29], the authors proposed a domain adaptation method that maps both the original and target data into a common domain, while keeping the topology of the input data probability distributions. For this purpose, the authors used a local dataset, where the sEMG data was acquired by repetitive gripping tasks while data was collected from 8 subjects. In addition to the above, Transfer Learning (TL) was also used to adopt a pre-trained model and leverage the knowledge acquired from multiple subjects and speed up the training process for the new users. In [30], [31], the authors proposed a TL-based algorithm adopting Convolutional Neural Networks (CNN) to transfer knowledge across multiple subjects for sEMG-based hand gesture recognition. The authors in [30], [31], applied the Myo armband to collect sEMG signals and used the fifth Niapro database, which contains data from 10 intact-limb subjects. The pre-training for each participant was done employing the training sets of the remaining nine participants and the average accuracy was obtained over the 10 participants of the Niapro DB5 [6]. Finally, References [32], [33] applied deep learning along with domain adaptation techniques for inter-session classification to improve the robustness for the long-term uses. Due to the variability of the signal space, the generalizability of existing techniques is questionable and it is not clear how they would perform in real-life scenarios when the training data is limited and extensive data collection cannot be done with high frequency to capture the changes.

In summary, there is an urgent need to develop adaptive learning methods with the focus on designing a classifier which can be adopted for new subjects based on only a few examples through a fast learning approach. This is a challenging task since many factors, such as electrode location and muscle fiber lengthening/shortening, can affect the collected sEMG signals. Moreover, the differences between users and the changes caused by amputations result in discrepancies between different conditions [2], [8]. To the best of our knowledge, this is the first time that *Few-shot Learning* is adopted in the literature to classify 49 hand gestures on *new* subjects using only a small (one to five) number of training examples.

III. MATERIAL AND METHODS

A. Database

The proposed FS-HGR architecture was evaluated on the Niapro [23]–[25] benchmark database, which is a publicly available dataset for hand gesture recognition tasks. Niapro is a widely used benchmark for evaluation of different models developed using sparse multichannel sEMG signals.

In this work, the second Niapro database [23] referred to as the DB2 was utilized. Delsys Trigno Wireless EMG system with 12 wireless electrodes (channels) was used in the DB2 dataset to collect electrical activities of muscles at a rate of

2 kHz. The dataset consists of signals collected from 28 men and 12 women with age 29.9 ± 3.9 years, among whom 34 are right-handed and 6 are left-handed. The DB2 consists of 50 gestures including wrist, hand, grasping, and functional movements along with force patterns from 40 healthy (intact-limb) subjects. The subjects repeated each movement 6 times, each time lasted for 5 seconds followed by 3 seconds of rest. More detail on the Ninapro database are described in Reference [23].

B. Pre-processing Step

Following the pre-processing procedure established in previous studies [16], [21]–[23], we used a 1st order low-pass Butterworth filter to smooth the electrical activities of muscles. Moreover, we applied μ -law transformation to magnify the output of sensors with small magnitude (in a logarithmic fashion), while keeping the scale of those sensors having larger values over time. This transformation approach has been used traditionally in speech and communication domains for quantization purposes. We propose to use it for scaling the sEMG signals as a pre-processing approach. The μ -law transformation was performed based on the following formulation

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

where $t \geq 1$ is the time index; x_t denotes the input to be scaled, and the parameter μ defines the new range. Here $\mu = 2,048$ was utilized, i.e., the scaled data points were distributed between 0 and 2,048. Afterwards, we fed the scaled sEMG signals to Minmax normalization. We empirically observed that the normalization of the scaled sEMG signals is significantly better than non-scaled sEMG signals.

This completes a brief introduction of the utilized dataset and the introduced pre-processing step. Next, we develop the proposed Meta Learning-based FS-HGR framework.

IV. THE FS-HGR FRAMEWORK

Meta-learning can be formalized as a sequence-to-sequence learning problem. The bottleneck is in the meta-learner's ability to internalize and refer to experience. To address this shortcoming for the gesture recognition task based on sparse multichannel sEMG, inspired by [26], we proposed a class of model architectures by combining temporal convolutions with attention mechanisms to enable the meta-learner to aggregate contextual information from experience. This integrated architecture allows the meta-learner to pinpoint specific pieces of information within its available set of inputs. Our main goal is to construct and train a hand gesture recognition model that can achieve rapid adaptation. Next, we first elaborate on the meta-learning concept.

A. The Meta-Learning Problem

A supervised learning task starts with a given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_D}$, consisting of N_D observations, where the i^{th} observation is denoted by \mathbf{x}_i , for $(1 \leq i \leq N)$, with

its associated label denoted by y_i . The main objective is to learn a (possibly non-linear) function $f(\cdot)$ defined based on its underlying parameters θ that maps each observation \mathbf{x}_i to its corresponding label, $y_i = f(\mathbf{x}_i; \theta)$. In a supervised learning approach, the dataset is divided into: (a) The training data $\mathcal{D}_{\text{train}}$ used for learning the parameters θ of the model; (b) The validation data \mathcal{D}_{val} utilized for tuning the hyper-parameters of the model, and; (c) The test data $\mathcal{D}_{\text{test}}$ for model evaluation.

In this context, we focused on meta-supervised learning, where the goal is generalization across tasks rather than across data points. Therefore, instead of using the aforementioned conventional data subsets (Items (a)-(c) above), we have a meta-set denoted by \mathcal{D} , which in turn splits into meta-train $\mathcal{D}_{\text{meta-train}}$, meta-validation $\mathcal{D}_{\text{meta-val}}$, and meta-test $\mathcal{D}_{\text{meta-test}}$ sub-datasets. Furthermore, one needs to construct different tasks (as shown in Fig. 1) within each meta-dataset. Task $\mathcal{T}_j \in \mathcal{D}$ is episodic and is defined by two components, a training set $\mathcal{D}_j^{\text{train}}$ for learning and a testing set $\mathcal{D}_j^{\text{test}}$ for evaluation, i.e., $\mathcal{T}_j = (\mathcal{D}_j^{\text{train}}, \mathcal{D}_j^{\text{test}})$.

Within the context of meta-learning, our focus is specifically on few-shot learning (typically referred to as k -shot learning with k being a small integer), which is briefly described next. In a N -way k -shot classification, our goal is training on $\mathcal{D}_{\text{meta-train}}$, where the input is the training set $\mathcal{D}_j^{\text{train}}$ and, a test instance $\mathbf{x}_j^{\text{test}} \in \mathcal{D}_j^{\text{test}}$. To be more precise, $\mathcal{D}_j^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{k \times N}$, where N classes are sampled from the meta-train set, and then k examples are sampled from each of these classes. To make predictions about a new test data point, $\mathbf{x}_j^{\text{test}} \in \mathcal{D}_j^{\text{test}}$, we produce a mapping function $f(\cdot)$ that takes as input $\mathcal{D}_j^{\text{train}}$ and $\mathbf{x}_j^{\text{test}}$ to produce the label $\hat{y}_j^{\text{test}} = f(\mathcal{D}_j^{\text{train}}, \mathbf{x}_j^{\text{test}}; \theta)$. Hyper-parameter selection is performed by using $\mathcal{D}_{\text{meta-val}}$. Generalization performance of the meta-learner is then evaluated on the $\mathcal{D}_{\text{meta-test}}$ [27].

Fig. 1 shows a $N = 5$ -way $k = 1$ -shot classification task, where inside each purple box is a separate dataset \mathcal{T}_j consisting of the training set $\mathcal{D}_j^{\text{train}}$ (on the Left-Hand Side (LHS) of the dashed line) and the $\mathcal{D}_j^{\text{test}}$ (on the Right-Hand Side (RHS) of the dashed line). In the illustrative example of Fig. 1, we are considering a 5-way 1-shot classification task where for each dataset, we have one example from each of the 5 classes (each given a label 1 to 5) in the training set and 1 example for evaluation from the test set of that specific task.

B. Description of the FS-HGR Model

In few-shot classification, the goal is to reduce the prediction error on data samples with unknown labels given a small training set. Inspired by [26], the proposed FS-HGR network receives as input a sequence of example-label pairs $\mathcal{D}_j^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{k \times N}$, followed by $\mathcal{D}_j^{\text{test}}$, which consists of an unlabelled example. The meta-learning model predicts the label of the final example based on the previous labels that it has seen. During the training phase, first, we sample N classes, with k examples per $\mathcal{D}_j^{\text{train}}$ (in terms of our running illustrative example, for each task, we have $k = 1$ sample from each of the underlying $N = 5$ classes). For constructing the $\mathcal{D}_j^{\text{test}}$, we sample an extra example from one of those selected classes. Afterwards, each set of the observations and labels

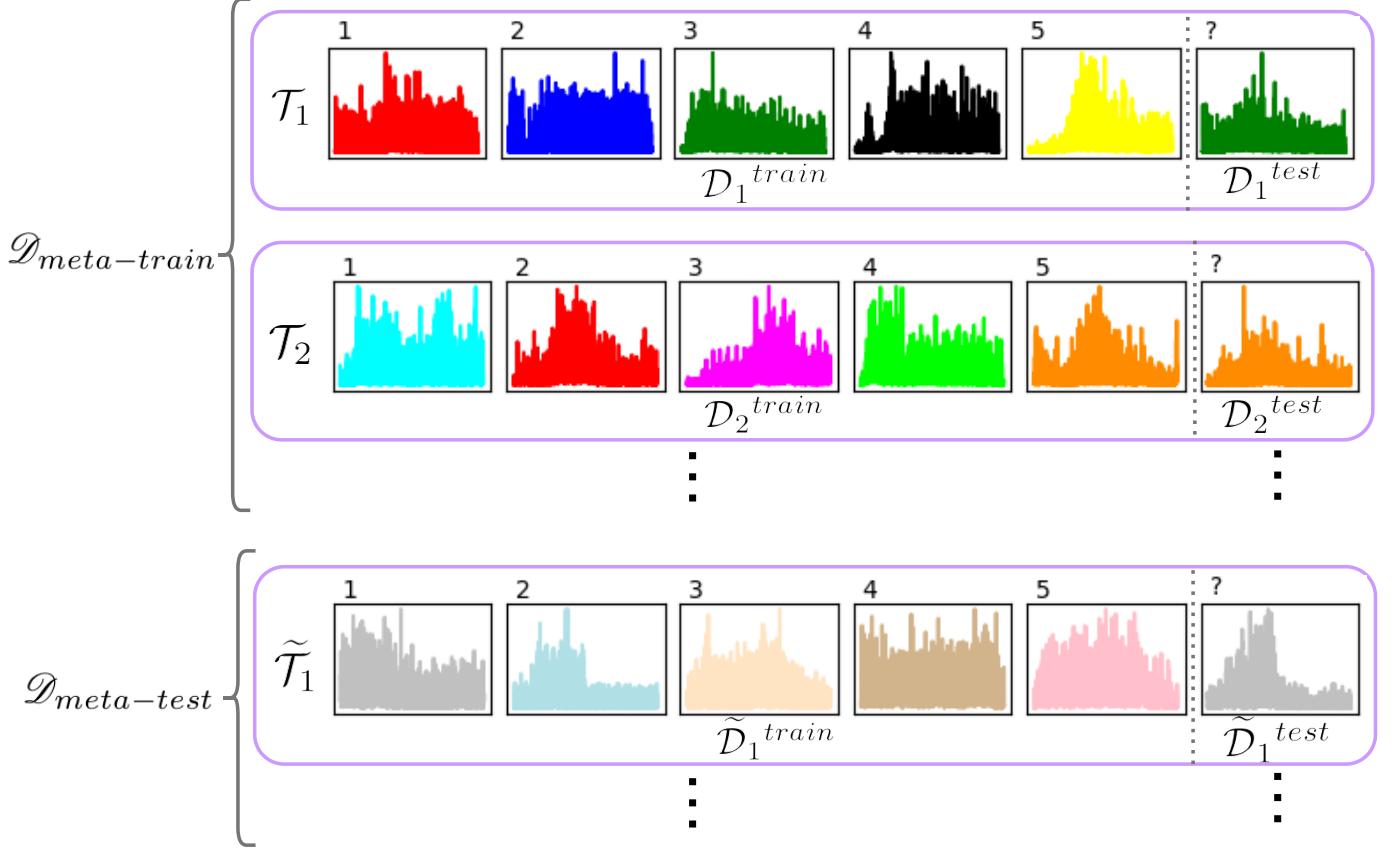


Fig. 1: 5-way 1-shot classification. Each task \mathcal{T} , represented in a purple box, is associated with a training set \mathcal{D}^{train} and a prediction set \mathcal{D}^{test} . Here, for constructing \mathcal{D}^{train} , first, 5 classes are sampled from the $\mathcal{D}_{meta-train}$, and then one example from each of these 5 classes (each corresponding with a label 1-5) are sampled. \mathcal{D}^{test} consists of 1 example sampled from one of those 5 classes. The $\mathcal{D}_{meta-test}$ is represented in the same approach, covering a different set of datasets which do not include any classes presented in any of the datasets in $\mathcal{D}_{meta-train}$. Moreover, $\mathcal{D}_{meta-val}$ is defined in the same way to determine the hyper-parameters of the model.

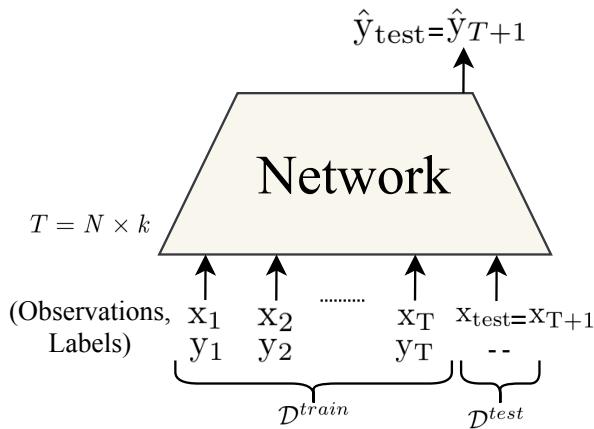


Fig. 2: For each task \mathcal{T}_j , the set of observations and labels are concatenated together and sequentially fed to the model. The final example is concatenated with a null label instead of True label. The network is supposed to predict the missing label of final example based the previous labels that it has seen. In N -way k -shot classification, N shows the number of classes which are sampled from whole set of labels, and k shows the examples that are sampled from each of those N classes.

are concatenated together (the final example is concatenated with a null label instead of the ground truth label as it is used for evaluation purposes), and then all $(N \times k + 1)$

Algorithm 1 THE TRAINING PROCEDURE

Input: $\mathcal{D}_{meta-train}$, and; mapping function $f(\cdot)$ with parameters θ .

Require. $p(\mathcal{T})$: distribution over tasks

- 1: **while** not done **do**
- 2: Sample batch of tasks $\mathcal{T}_j \sim p(\mathcal{T})$
- 3: **for all** \mathcal{T}_j **do**
- 4: Split \mathcal{T}_j into \mathcal{D}_j^{train} and \mathcal{D}_j^{test}
- 5: Predict the missing label of final example of \mathcal{T}_j :

$$\hat{y}_{test} = f(\mathcal{D}_j^{train}, \mathcal{D}_j^{test}; \theta)$$
- 6: **end for**
- 7: Update θ using $\sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(\hat{y}_{test}, y_{test})$
- 8: **end while**

are sequentially fed to the network. Finally, the loss \mathcal{L}_j is computed between the predicted and ground truth label of the $(N \times k + 1)^{th}$ example. During such a training mechanism, the network learns how to encode the first $N \times k$ examples to make a prediction about the final example [26]. The training procedure is described in Algorithm 1 and the schematic of the model is shown in Fig. 2 (further information is available at the link provided in Reference [38]).

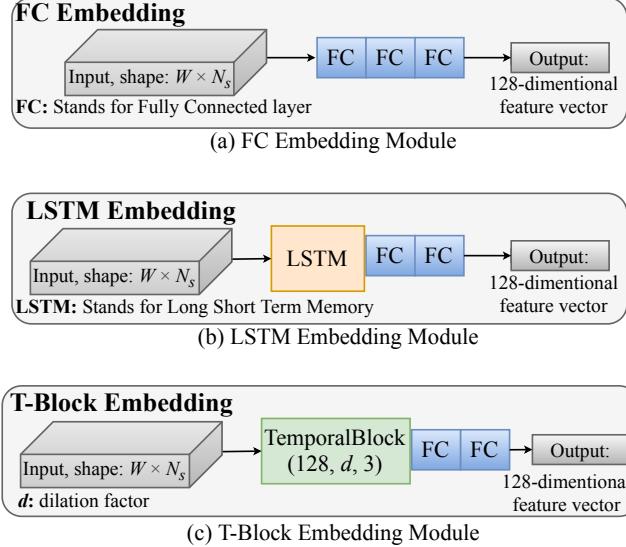


Fig. 3: **The Embedding Module**, which converts an input with size $(W \times N_S)$, W stands the sequence length and N_S shows the number of input features, to a 128-dimensional feature vector. (a) **FC Embedding Module**, which uses three FC layers to outputs a 128-dimensional feature vector. (b) **LSTM Embedding Module**, which adopts a LSTM layer followed by two FC layers. (c) **T-Block Embedding Module**, which consists of a Temporal Block with number of filters $f = 128$, kernel size $k = 3$, and dilation factor d , followed by two FC layers.

C. The Building Modules of the FS-HGR Framework

After completion of the pre-processing step, sEMG signals acquired from N_S number of sensors are segmented by a window of length of $W = 200$ ms selected to satisfy the acceptable delay time [34], i.e., the window length W is required to be under 300 ms. Finally, sliding window with steps of 50 ms is considered for segmentation of the sEMG signals.

1) **The Embedding Module:** To develop the FS-HGR for few-shot learning, we aimed to first extract a 128-dimensional feature vector from each observation with size of $(W \times N_S)$, where W stands for the sequence length and N_S shows the number of input features, e.g., in the experiments N_S is equal to 12 as twelve sensing channels are available. The “Embedding Module” is, therefore, used to extract a 128-dimensional feature vector, which is then provided as input to the proceeding modules within the proposed architecture.

Adopting a proper Embedding Module has a significant effect on the results. For validating our claim, therefore, we utilized four different Embedding Modules:

- The first Embedding Module, referred to as the *FC Embedding*, consists of three Fully Connected (FC) layers to output a 128-dimensional feature vector from each observation. The first FC layer in the FC Embedding Module is used to increase the input dimensional to $(W \times 128)$. Subsequently, the second (which is followed by ReLU activation function) and third FC layers with output size of 100 and 1, respectively, are adopted to reduce the sequence length of each observation to (1×128) (Fig. 3(a));
- LSTM Embedding:* Fig. 3(b) illustrates the second Embedding Module, referred to as the *LSTM Embedding*,

which utilizes a Long Short-Term Memory (LSTM) layer as its first block followed by two FC layers. The LSTM layer takes the observation with input size 12 and converts it to an output with 128 features. Then, the two FC layers are adopted to reduce the observation’s sequence length to 1;

- T-Block Embedding I:* This third Embedding Module utilizes the *TemporalBlock Module* (which will be described in next sub-section) consisting of $f = 128$ 1D-Convolutions with kernel size $k = 3$, and dilation factor $d = 1$ as its first block. The *TemporalBlock Module* is followed by two FC layers to decrease the input’s sequence length to 1 as shown in Fig. 3(c), and;
- T-Block Embedding II:* This embedding is similar in nature to the one described above in Item (iii), however, here the goal is to examine the effect of increasing the size of the receptive field. As such, the fourth Embedding Module utilizes two *TemporalBlock Modules* with $d = 1$ and $d = 2$. It is noteworthy to mention that the first FC layer in both LSTM and T-Block Embedding modules are followed by ReLU activation function.

2) **The TemporalBlock Module:** Inspired by [10], [12], [26], [35], [36], the proposed FS-HGR few-shot learning architecture utilizes *Dilated Causal 1D-Convolutions* over the temporal dimension. The proposed architecture, therefore, provides several advantages over RNNs such as low memory requirement and faster training. In addition, and unlike conventional CNNs, by incorporation of dilated causal convolutions, we increased the receptive field of the network and as such benefit from the time-series nature of the input.

As shown in Fig. 4(a), each *TemporalBlock* consists of two dilated causal 1D-convolutions, each with dilation factor d , filter size k , and f number of filters. To learn the complex structure of the underlying data, each Dilated Causal 1D-Convolutions is followed by a ReLU activation function. Finally, by concatenating the results and the input, the training speed can be considerably improved. This module takes an input with size $(C_{in} \times l)$ and output a tensor with size $(C_{out} \times l)$. Here, l denotes the sequence length and is equal to $(N \times k + 1)$, which N shows the number of class samples from the whole set of labels, and k shows the number of examples per each class.

3) **The TemporalConvNet Module:** The benefit that comes with the designed “TemporalConvNet” module is that its training procedure is much faster and efficient compared to LSTM or Gated Recurrent Unit (GRU) architectures. In other words, through this approach one complete sequence can be processed through only one forward pass, while in RNN-based models this, typically, needs several passes due to temporally linear hidden state dependency. The TemporalConvNet module consists of a series of TemporalBlock modules with exponentially growing dilation factors d . More specifically, as shown in Fig. 4(b), for an input with sequence length $l = (N \times k + 1)$, the TemporalConvNet consists of $Z = \lceil \log_2 l \rceil$ number of TemporalBlock modules. The dilation factors d for the TemporalBlock modules are equal to $[1, 2, 4, \dots, 2^{Z-1}]$, respectively.

4) **The Attention Module:** The final constituent module within the proposed FS-HGR architecture is referred to as the

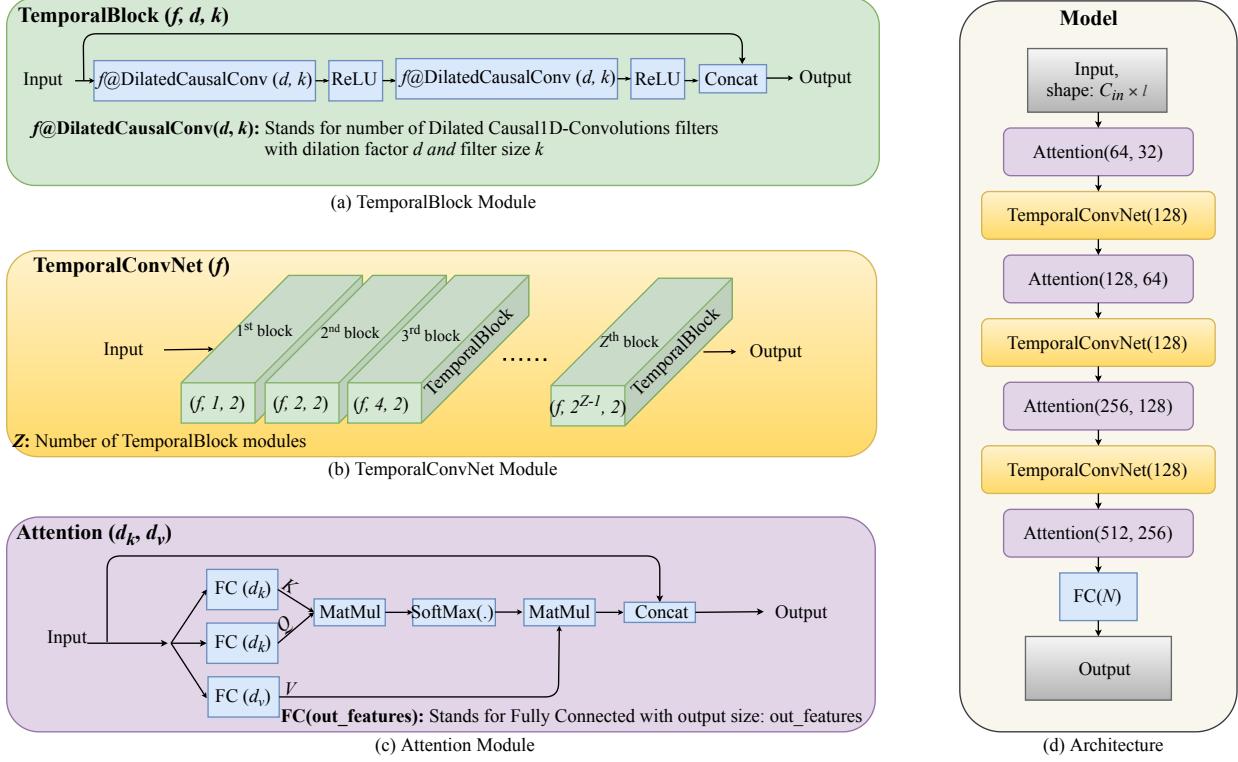


Fig. 4: (a) **The TemporalBlock Module**, which consists of f Dilated Causal 1D-Convolutions with dilation factor d and kernel size k . This module converts an input with C_{in} features to an output with C_{out} features. The sequence length of the input l is equal to $(N \times k + 1)$, where N shows the number of classes and k denotes the number of examples of each class. (b) **The TemporalConvNet Module**, which consists of a series of TemporalBlock modules (green ones). The kernel size of each TemporalBlock Module k is equal to 2; however, their dilation factor d increases exponentially. (c) **The Attention Module**, which consists of three FC layers with output size d_k , d_k , and d_v , respectively, to produce matrix Q , K , and V . (d) **The Architecture**, consisting of three TemporalConvNet modules (yellow ones), and four Attention modules (purple ones). Here, 128 denotes the number of filters f in Dilated 1D-Convolutions. The architecture is supposed to predict the missing label of the $(N \times k + 1)^{th}$ example in each task \mathcal{T}_j .

“Attention Module,” included with the objective of pinpointing a specific type of information within the available (possibly significantly large) context [37]. Attention mechanism has been recently utilized [13] within the context of sEMG-based hand gesture recognition, where the experiments showed attention’s capability to learn a time-domain representation of multichannel sEMG data. By integrating the TemporalConvNet, described above, and the Attention Module, essentially we provided the FS-HGR architecture with the capability to access the past experience without any limitations on the size of experience that can be used effectively. Furthermore, in the FS-HGR framework we used the Attention Module at different stages to provide the model with the ability to learn how to identify and select pieces of useful information and its appropriate representation from its experience.

As shown in Fig.4(c), to get queries, keys, and values, three linear transformations are applied to the input. The attention mechanism then compares queries to each of the key values with a dot-product, scaled by $\sqrt{d_k}$, which results compatibility scores. To obtain attention distribution over the values, softmax function is applied to the scores. Then, we computed the weighted average of the values, weighted by the attention distribution. In practice, the keys, values, and queries are packed together into matrices K , V , and Q , respectively. The

matrix of outputs is obtained as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad (2)$$

where d_k stands for length of the key vector in matrix \mathbf{K} . Then, the results and inputs are concatenated together. This completes description of the modules incorporated to construct the proposed FS-HGR framework. Next, we present its overall architecture.

D. The Architecture

The overall structure of the proposed FS-HGR architecture consists of four Attention modules, where the first three ones are followed by a TemporalConvNet module. The final Attention module is followed by a FC layer to produce the label of the final example in each task \mathcal{T}_j . More specifically, after feeding each observation with size $W \times N_S$ to an Embedding Module, we obtained a 128-dimensional feature vector (Fig. 3). Then, for constructing each task \mathcal{T}_j with sequence length l (Fig. 2), the set of observations (each observation is converted to a 128-dimensional feature vector) and labels are concatenated. The final observation in the sequence is concatenated with a null label instead of a True label. The network is supposed to predict the missing label of the final example based on the previous labels that it has seen. In

TABLE I: Experiment 1: 5-way, 1-shot, 5-shot, and 10-shot classification accuracies on *new repetitions with few-shot observation*. The classification on new repetitions with few-shot observation are performed by using Meta-supervised Learning approach. This table also shows a comparison between our methodology (Meta-supervised) learning and previous works where Supervised learning methodology is used.

Meta-Supervised Learning Proposed Method	The Embedding Module	5-way Accuracy		
		1-shot	5-shot	10-shot
	FC Embedding	72.59%	85.13%	89.26%
	LSTM Embedding	75.03%	84.06%	88.45%
	T-Block Embedding I	73.46%	85.94%	89.40%
	T-Block Embedding II	74.89%	85.88%	89.70%
Supervised Learning Previous Works	Previous Works	Accuracy		
	Wei <i>et al.</i> [17]	83.70%		
	Hu <i>et al.</i> [18]	82.20%		
	Ding <i>et al.</i> [19]	78.86%		
	Zhai <i>et al.</i> [20]	78.71%		
	Geng <i>et al.</i> [21]	77.80%		
	Atzori <i>et al.</i> [22]	75.27%		

summary, to perform the hand gesture recognition task, the FS-HGR framework is constructed based on different modules as shown in Fig. 4(d).

V. EXPERIMENTS AND RESULTS

In this section, we describe a comprehensive set of experiments to analyse and evaluate the proposed FS-HGR framework. It is worth mentioning that in few-shot classification, we would like to classify inputs in N classes when we have just k examples per class. To evaluate the proposed architecture for N -way k -shot classification, we randomly sampled N classes from the overall classes, and then sampled k examples from each class. Then, we fed the $(N \times k)$ observation-label pairs to the model followed by a new unlabelled example sampled from one of the N classes. The objective of the FS-HGR model is to predict the missing label of the $(N \times k + 1)^{th}$ in the sequence.

In the following, we present three evaluation scenarios. In all experiments, Adam optimizer was used for training purposes with learning rate of 0.0001. Different models were trained with a mini-batch size of 64 except in 10-way 5-shot classification where mini-batch size of 32 was used. For measuring the classification performance, the loss \mathcal{L}_j was computed between the predicted and ground truth label of $(N \times k + 1)^{th}$ example in each task \mathcal{T}_j . The average loss was computed using Cross-entropy loss. Finally, the average accuracy is reported on the $(N \times k + 1)^{th}$ example.

TABLE II: Experiment 2(a): 5-way and 10-way, 1-shot and 5-shot classification accuracies based on *new subjects with few-shot observation*. In this experiment, we adopted four different Embedding Modules: (i) FC Embedding; (ii) LSTM Embedding; (iii) T-Block Embedding I, and; (iv) T-Block Embedding II.

The Embedding Module	5-way Accuracy		10-way Accuracy	
	1-shot	5-shot	1-shot	5-shot
FC Embedding	62.87%	78.90%	43.47%	68.59%
LSTM Embedding	64.46%	79.82%	49.58%	69.93%
T-Block Embedding I	67.81%	81.08%	50.31%	69.94%
T-Block Embedding II	66.98%	81.29%	52.05%	70.71%

Experiment 1: Classification on New-Repetitions with Few-Shot Observation. The first experiment shows that our proposed network is applicable when we had new repetitions with few-shot observation on the target. We evaluated our proposed architecture when $\mathcal{D}_{meta-train}$ consisted of the 2/3 of the gesture trials of each subject (following Reference [22], repetitions 1, 3, 4, and 6 repetitions were used for training purposes), and $\mathcal{D}_{meta-test}$ consisted of the remaining repetitions. Table I shows our results when using few-shot classification as well as previous works which used supervised learning. From Table I, it can be observed that the proposed FS-HGR architecture outperformed existing methodologies when evaluated based on the same setting, i.e., 85.94% best accuracy with the FS-HGR compared to 83.70% best accuracy achieved by the state-of-the-art. Although this improvement is relatively small, the following Experiments 2 and 3 provide further evidence for the superior performance of the proposed approach.

Experiments 2: Classification on New-Subject with Few-Shot Observation. In this scenario, like the previous experiment, the second Ninapro database DB2 was utilized. It consists of 49 gestures plus rest from 40 intact-limb subjects. In this experiment, to validate our claim that the proposed FS-HGR architecture can classify hand gestures of new subjects just by training with a few examples, we split the DB2 database into $\mathcal{D}_{meta-train}$, $\mathcal{D}_{meta-val}$, and $\mathcal{D}_{meta-test}$ such that the subjects in these meta-sets are completely different (i.e., there is no overlap between the meta-sets). In other words, $\mathcal{D}_{meta-train}$ consists of the first 27 subjects, while $\mathcal{D}_{meta-val}$ includes the sEMG signals from the 28th subject to 32^{ed} subject (5 subjects). Finally, we evaluated our model on the remaining subjects, i.e., $\mathcal{D}_{meta-test}$ consists of the final 8 subjects in the DB2 database.

It is noteworthy to mention that the proposed network is trained once and shared across all participants (which is different from previous works that trained the model separately for each participant). For constructing task \mathcal{T}_j , however, we can feed data in two different approaches:

- **Experiment 2(a):** In the first approach, for constructing \mathcal{D}_j^{train} for each task \mathcal{T}_j , we sampled all of the N classes from a specific user, which was randomly selected from the existing participants. This is the more realistic and practical scenario.

TABLE III: Comparison of 5-way, 1-shot and 5-shot classification accuracies between the Experiment 2(a) and 2(b) based on *new subjects with few-shot observation*.

The Embedding Module	Experiment 2(a)		Experiment 2(b)	
	5-way Accuracy			
	1-shot	5-shot	1-shot	5-shot
FC Embedding	62.87%	78.90%	72.69%	86.08%
LSTM Embedding	64.46%	79.82%	75.56%	89.14%
T-Block Embedding I	67.81%	81.08%	75.11%	89.66%
T-Block Embedding II	66.98%	81.29%	77.08%	90.47%

- *Experiment 2(b)*: In the second approach, for constructing \mathcal{D}_j^{train} , N classes were sampled from different participants.

Table II shows few-shot classification accuracies for Experiment 2(a) based on four different embedding modules. The adaptive learning method of the proposed FS-HGR focuses on transfer learning information between a source and a target domain despite the existence of a distribution mismatch between $\mathcal{D}_{meta-train}$ and $\mathcal{D}_{meta-test}$. The results reported in Table II show that the proposed mechanism achieves acceptable results despite the fact that the sEMG signals are user-dependant. Table III shows a comparison of 5-way classification accuracies between Experiments 2(a) and 2(b). As was it expected, Experiment 2(b) achieved better results, which is due to the presence of variations among the probability distribution of sEMG signals obtained from different subjects. However, this is not a practical setting as in practice all of the N classes in \mathcal{D}_j^{train} comes from the same user (i.e., Experiment 2(a)). Experiment 2(a) is the more realistic and challenging one. Experiment 2(b) is included for completeness and comparison purposes.

Finally, it is worth noting that adoption of few-shot learning within the FS-HGR framework has resulted in significant reduction in the required training time for users. As explained before, the dataset was collected from 40 people including 49 gesture with 6 repetition of each, where each repetition lasted 5 seconds. In previous studies, 4 repetitions, 20 seconds in total, of each user's gestures were considered for the training purpose, and the remaining 2 repetitions were used for their model evaluation. Commonly, sliding-window with window size of 200 ms is leveraged for feeding data to the models. However, in our few-shot based model, we used 6 repetitions of 27 subjects for training, and the model did not see any data of the remaining subjects during the learning procedure. The gained experience during the training is leveraged to tune the model to a new user by seeing a small number of intervals (each of duration 200 ms). More specificity, for a new user in a N -way k -shot classification problem, we just used $N \times k$ windows. For example, in 5-way 1-shot, we just need 5 windows (1 second in total) to recalibrate the model. Here, unlike previous methods, we do not have to train the model from scratch for a new user, or fine tune model for 4 repartitions of new user's gestures. We needed only $N \times k$

TABLE IV: Experiment 3: 5-way, 1-shot, 5-shot, and 10-shot classification accuracies based on *new gesture with few-shot observation*.

The Embedding Module	5-way Accuracy		
	1-shot	5-shot	10-shot
FC Embedding	45.94%	67.20%	79.87%
LSTM Embedding	46.05%	71.76%	81.58%
T-Block Embedding I	49.78%	71.57%	83.41%
T-Block Embedding II	45.48%	73.36%	–

windows from a new user, to adapt the trained model for this new user. Therefore, by using few-shot learning, the proposed FS-HGR framework has the potential to significantly reduce the training time.

Experiment 3: Classification on New-Gestures with few-shot observations. In this scenario, the goal is evaluating the capability of the proposed FS-HGR architecture when the target consists of solely out-of-sample gestures (i.e., new gestures with few-shot observation). Performing well in this task allows the model to evaluate new observations, exactly one per novel hand gesture class. In this experiment, the Ninapro database DB2 was used. The DB2 dataset includes three sets of exercises denoted by Exercise *B*, *C*, and *D*. Exercise *B* includes 8 isometric and isotonic hand configurations and 9 basic movements of the wrist; Exercise *C* consists of 23 grasping and functional movements; and finally, Exercise *D* consists of 9 force patterns. For training purposes, $\mathcal{D}_{meta-train}$ consisted of the first 34 gestures of each user, which is equal to approximately 68% of the total gestures. $\mathcal{D}_{meta-val}$ included 6 gestures or 12% of the total gestures. The remaining gestures (9 gestures), were used in $\mathcal{D}_{meta-test}$ for evaluation purposes. Exercises *B* and *C* were, therefore, used for training and validation, and Exercises *D*, with different gestures, were used for test purposes. Table IV shows the efficiency of the proposed model when we had out-of-sample gestures in the target. The model predicted unknown class distributions in scenarios where few examples from the target distribution were available.

VI. CONCLUSION

We proposed a novel few-shot learning recognition approach for the task of hand gesture recognition via sEMG signals. The proposed FS-HGR framework could quickly generalize after seeing very few examples from each class. This is achieved by exploiting the knowledge gathered from previous experiences to accelerate the learning process performed by a new subject. The experience gained over several source subjects is leveraged to reduce the training time of a new target user. In this way the learning process does not start every time from the beginning, and instead refines. The ability to learn quickly based on a few examples is a key characteristic of the proposed FS-HGR framework that distinguishes this novel architecture from its previous counterparts. A second

contribution of the paper is its capability to address the user-dependent nature of the sEMG signals. The proposed FS-HGR framework transfers information between a source and a target domain despite the existence of a distribution mismatch among them. This would dramatically reduce the number of required cumbersome training sessions leading to a drastic reduction in functional prosthesis abandonment.

ACKNOWLEDGEMENT

This work was supported by Borealis AI through the Borealis AI Global Fellowship Award.

REFERENCES

- [1] N. Jiang, S. Dosen, K.R. Muller, D. Farina, "Myoelectric Control of Artificial Limbs- Is There a Need to Change Focus?" *IEEE Signal Process. Mag.*, vol. 29, pp. 150-152, 2012.
- [2] D. Farina, R. Merletti, R.M. Enoka, "The Extraction of Neural Strategies from the Surface EMG," *J. Appl. Physiol.*, vol. 96, pp. 1486-95, 2004.
- [3] D. Esposito, E. Andreozzi, G.D. Gargiulo, A. Fratini, G. D'Addio, G.R. Naik, and P. Bifulco, "A Piezoresistive Array Armband with Reduced Number of Sensors for Hand Gesture Recognition," *Frontiers in Neurorobotics*, vol. 13, p. 114, 2020.
- [4] M. Tavakoli, C. Benussi, P.A. Lopes, L.B. Osorio, and A.T. de Almeida, "Robust Hand Gesture Recognition with a Double Channel Surface EMG Wearable Armband and SVM Classifier," *Biomedical Signal Processing and Control*, vol. 46, pp. 121-130, 2018.
- [5] G.R. Naik, A.H. Al-Timemy, H.T. Nguyen, "Transradial Amputee Gesture Classification using an Optimal Number of sEMG Sensors: an Approach using ICA Clustering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 8, pp. 837-846, 2015.
- [6] S. Pizzolato, L. Tagliapietra, M. Cognolato, M. Reggiani, H. Muller, and M. Atzori, "Comparison of Six Electromyography Acquisition Setups on Hand Movement Classification Tasks," *PLoS ONE*, vol. 12, no. 10, pp. 1-7, 2017.
- [7] C. Castellini, P. Artemiadis, M. Wininger, A. Ajoudani, M. Alimusaj, A. Bicchi, B. Caputo, W. Craelius, S. Dosen, K. Englehart, and D. Farina "Proceedings of the First Workshop on Peripheral Machine Interfaces: Going Beyond Traditional Surface Electromyography," *Frontiers in neurorobotics*, 8, p. 22, 2014.
- [8] D. Farina, N. Jiang, H. Rehbaum, A. Holobar, B. Graumann, H. Dietl, and O. C. Aszmann, "The Extraction of Neural Information from the Surface EMG for the Control of Upper-limb Prostheses: Emerging Avenues and Challenges..," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no.4, pp. 797-809, 2014.
- [9] A. K. Clarke *et al.*, "Deep Learning for Robust Decomposition of High-Density Surface EMG Signals," *IEEE Transactions on Biomedical Engineering*, 2020, In Press.
- [10] E. Rahimian, S. Zabihi, S. F. Atashzar, A. Asif, and A. Mohammadi, "Surface EMG-Based Hand Gesture Recognition via Hybrid and Dilated Deep Neural Network Architectures for Neurorobotic Prostheses," *Journal of Medical Robotics Research*, 2020, pp. 1-12.
- [11] E. Rahimian, S. Zabihi, F. Atashzar, A. Asif, A. Mohammadi, "XceptionTime: Independent Time-Window XceptionTime Architecture for Hand Gesture Classification," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.
- [12] E. Rahimian, S. Zabihi, S. F. Atashzar, A. Asif, and A. Mohammadi, "Semg-based Hand Gesture Recognition via Dilated Convolutional Neural Networks," *Global Conference on Signal and Information Processing, GlobalSIP*, 2019.
- [13] D. Josephs, C. Drake, A. Heroy, and J. Santerre, "sEMG Gesture Recognition with a Simple Model of Attention," *arXiv preprint arXiv:2006.03645*, 2020.
- [14] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, "Hand Gesture Recognition using Compact CNN via Surface Electromyography Signals," *Sensors*, vol. 20, no.3, p. 672, 2020.
- [15] Y. Peng, H. Tao, W. Li, H. Yuan and T. Li, "Dynamic Gesture Recognition based on Feature Fusion Network and Variant ConvLSTM," *IET Image Processing*, vol. 14, no. 11, pp. 2480-2486, 18 9 2020.
- [16] W. Wei, Y. Wong, Y. Du, Y. Hu, M. Kankanhalli, and W. Geng, "A Multi-stream Convolutional Neural Network for sEMG-based Gesture Recognition in Muscle-computer Interface," *Pattern Recognition Letters*, 119, pp. 131-138, 2019.
- [17] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli, and W. Geng, "Surface Electromyography-based Gesture Recognition by Multi-view Deep Learning," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2964-2973, 2019.
- [18] 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* 13, no. 10, 2018.
- [19] Z. Ding, C. Yang, Z. Tian, C. Yi, Y. Fu, and F. Jiang, "sEMG-based Gesture Recognition with Convolution Neural Networks," *Sustainability* 10, no. 6, p. 1865, 2018.
- [20] X. Zhai, B. Jelfs, R. H. Chan, and C. Tin, "Self-recalibrating Surface EMG Pattern Recognition for Neuroprosthesis Control based on Convolutional Neural Network," *Frontiers in neuroscience*, 11, p.379, 2017.
- [21] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, "Gesture Recognition by Instantaneous Surface EMG Images," *Scientific reports*, 6, p. 36571, 2016.
- [22] M. Atzori, M. Cognolato, and H. Müller, "Deep Learning with Convolutional Neural Networks Applied to Electromyography Data: A Resource for the Classification of Movements for Prosthetic Hands," *Frontiers in neurorobotics* 10, p.9, 2016.
- [23] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.G.M Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography Data for Non-Invasive Naturally-Controlled Robotic Hand Prostheses," *Scientific data* 1, no. 1, pp. 1-13, 2014.
- [24] A. Gijsberts, M. Atzori, C. Castellini, H. Müller, and B. Caputo, "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, 2014.
- [25] M. Atzori, A. Gijsberts, I. Kuzborskij, S. Heynen, A.G.M Hager, O. Deriaz, C. Castellini, H. Müller, and B. Caputo, "A Benchmark Database for Myoelectric Movement Classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2013.
- [26] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A Simple Neural Attentive Meta-Learner," *arXiv preprint arXiv:1707.03141*, 2017.
- [27] S. Ravi, and H. Larochelle, "Optimization as a Model for Few-shot Learning," *International Conference on Learning Representations (ICLR)*, 2016.
- [28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic Meta-learning for Fast Adaptation of Deep Networks," *In Proceedings of the 34th International Conference on Machine Learning-Volume 70*, vol. 100, pp. 1126-1135, 2017.
- [29] R. Chattopadhyay, N. C. Krishnan, and S. Panchanathan, "Topology Preserving Domain Adaptation for Addressing Subject based Variability in SEMG Signal," in *Proc. AAAI Spring Symp., Comput. Physiol.*, 2011, pp. 4-9.
- [30] U. Côté-Allard, C.L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, and B. Gosselin, "Deep Learning for Electromyographic Hand Gesture Signal Classification using Transfer Learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 760-771, 2019.
- [31] 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," *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2017, pp. 1663–1668.
- [32] M. Zia ur Rehman, A. Waris, S.O. Gilani, M. Jochumsen, I.K. Niazi, M. Jamil, D. Farina, and E.N. Kamavuako, "Multiday EMG-based Classification of Hand Motions with Deep Learning Techniques," *Sensors*, vol. 18, no. 8, p. 2497, 2018.
- [33] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, "Surface Emg-based Inter-session Gesture Recognition Enhanced by Deep Domain Adaptation," *Sensors*, vol. 17, no. 3, p. 458, 2017.
- [34] B. Hudgins, P. Parker, and R.N. Scott, "A New Strategy for Multifunction Myoelectric Control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, p.82-94, 1993.
- [35] S. Bai, J.Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv preprint arXiv:1803.01271*,2018.
- [36] A.V.D. Oord, *et al.*, "Wavenet: A Generative Model for Raw Audio," *ArXiv preprint arXiv:1609.03499*, 2016.
- [37] A. Vaswani, N. Shazeer, J. Uszkoreit, L. Jones, A. Gomez N., L. Kaiser, and I. Polosukhin, "Attention is All You Need," *arXiv preprint arXiv:1706.03762*, 2017a.
- [38] <https://ellarahimian.github.io/FS-HGR/>