Toward Deep Generalization of Peripheral EMG-Based Human-Robot Interfacing: A Hybrid Explainable Solution for NeuroRobotic Systems

Paras Gulati, Qin Hu, *Student Member, IEEE*, and S. Farokh Atashzar, *Member, IEEE*

***Abstract*—This letter investigates the feasibility of a generalizable solution for human-robot interfaces through peripheral multichannel Electromyography (EMG) recording. We propose a tangential approach in comparison to the literature to minimize the need for (re)calibration of the system for new users. The proposed algorithm decodes the signal space and detects the common underlying *global* neurophysiological components, which can be detected robustly across various users, minimizing the need for retraining and (re)calibration. The research question is how to go beyond techniques that detect a high number of gestures for a given individual (which requires extensive calibration) and achieve an algorithm that can detect a lower number of classes but without the need for (re)calibration. The outcomes of this letter address a challenge affecting the usability and acceptance of advanced myoelectric prostheses. For this, the paper proposes an explainable generalizable hybrid deep learning architecture that incorporates CNN and LSTM. We also utilize the GradCAM analysis to explain and optimize the structure of the generalized model, securing higher computational performance whiles proposing a shallower design.**

***Index Terms*—Electromyography, machine learning, medical robotics, prosthetics.**

I. Introduction

PERFORMANCE and efficacy of any neurorobotic systems, including robotic prostheses and neurorehabilitation robotic systems, depend critically on the quality of interfacing with human biomechanics and, more importantly, human neural systems. In this regard, it is known that in the context of neurorobotic prostheses, errors in detecting the intended gesture of an individual with the lack of a biological limb can significantly increase the mental and physical load and results in a high rejection rate of the technology. Another contributing problem to the current high rejection rate is the need for extensive calibration, which may take even a full-day of visit and stay in a clinic. Thus, although peripheral EMG-based robotic prosthesis has been investigated widely in the literature [1], [2], the current commercialized examples have very limited performance. Regarding neurorehabilitation robotic systems [3]–[6], which have been investigated in the literature for helping with the recovery process of neurologically-damaged patients, it is known that a robot which does not properly respond to the intended motion of a user will have limited performance in terms of recovery due to reduced engagement and participation of the user in robotic rehabilitation procedures (see our recent literature review [3], [7]). Motivated by the above-mentioned notes, processing of non-invasive surface electromyography (sEMG) signals for detection of the intended gestures of a human user (in particular amputees) has a long history [8], and there is extensive research regarding the performance of classical and more recently advanced machine learning algorithms for addressing this need [9] [10]. An ideal human-robot interface can detect and decode the intended motions of the user with the high spatiotemporal resoltion, using minimum calibration and recalibration. Conventional research on the topic of EMG-based human-robot interfaces has been focused on extracting spectral or temporal features from EMG signals to be used via classic machine learning classification algorithms to detect the intended gesture of the user for control of prostheses. More recently, deep learning algorithms have been utilized and tested [11] on major publicly available datasets (such as NinaPro [12]) to maximize the accuracy of the system for a very large number of gestures (for example 17 and more) relying on extensive user-specific data collection for calibration and training. High accuracy has been reported (please see our recent work on this topic [13] and other relevant publications [14]). Thanks to the wide availability of standard datasets during the last two years, the performance of these systems has been improved and compared systematically, resulting in very high accuracy of gesture detection even for a very high number of classes (>17). Recurrent Neural Networks (RNN) [15] and Convolution Neural Networks (CNNs) [16] have been used in the literature for this purpose [17]. RNN has been seen as the default choice for dealing with sequential data and has been used extensively in EMG classification [18] [19]. However, vanilla RNN suffers from vanishing gradient problems. Vanishing gradient is a problem where the gradients

Manuscript received October 15, 2020; accepted February 5, 2021. Date of publication February 25, 2021; date of current version March 17, 2021. This letter was recommended for publication by Associate Editor D. Oetomo and Editor P. Valdastri upon evaluation of the reviewers' comments. This work was supported by the National Science Foundation (Awards: # 2037878, and # 2031594). *(Corresponding author: S. Farokh Atashzar.)*

Paras Gulati and Qin Hu are with the Department of Electrical and Computer Engineering, New York University (NYU), Brooklyn, NY 11201 USA (e-mail: paras.gulati@nyu.edu; qh503@nyu.edu).

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

Digital Object Identifier 10.1109/LRA.2021.3062320

2377-3766 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

do not change much in deeper layers of the network, and the Gradient Descent algorithm manages to change only the gradients of layers closer to the output layer. A better suited recurrent network variant is the Long Short Term Memory (LSTM) network [20]. In addition to the above, CNNs are also being used in sequential data classification, and thus, many researchers have successfully leveraged CNNs for EMG classification [21] [22].

Despite the high performance of the deep neural network, the main concern is the need for significant calibration and the sensitivity to changes and minimum generalizability for EMG classification. It means that for every new user, there is a need for an extended calibration, which sometime may take more than a day. However, the neurophysiological nature of EMG is quite variable, which will restrict the usability of such an approach. To address the mentioned concern, in this letter, we propose a different vision regarding the use of deep neural networks for decoding human intention through the processing of the multichannel electromyography data. Here we propose to train on the system with a lower number of gestures (for example, four frequently used gestures during activities of daily living); however, challenge the system to detect the gestures correctly across subjects. This will have an imperative application in human-robot interfacing as it addresses a major need for calibration.

In this letter, to achieve the proposed goal, besides the mentioned training methodology, we utilize a hybrid approach combining LSTM and Dilated-CNN. The model showed a powerful performance for generalization in this work. We use causal CNN to maintain causality and avoid the information leakage from the future to the past. Moreover, we use dilated kernels [23] with the 1-Dimensional CNN layers to capture generalizable features over a broader temporal range taking into account long- and short-term histories. We first evaluate the performance of the system to address the classical problem of gesture detection in a user-specific manner for a high number of classes, and then we will conduct a comprehensive generalizability analysis.

It should be noted that deep learning methods are conventionally treated as a black-box system. Using such an approach, we can only see the input and output of the deep learning model but hardly comprehend why the neural network reached a particular decision. It is imperative to know what the distinguishing patterns in the data are based on which the neural network is able to identify the bases of its decision making. In this work, we use the GradCAM analysis [24] method to investigate how the activations in various layers of the network are contributing toward the decision by extracting certain neurophysiological features. We will use the results of the analysis to optimize the network and remove the layers which do not contribute to the generalizable module. We will show that through GradCAM analysis, we can detect the least contributing layers, optimizing the design and size of the network, maximizing practicality, and preserving the performance in terms of gesture prediction.

II. Database

NinaPro data set has been used during the last five years as the benchmark for evaluating the functionality of various machine learning algorithms applied to EMG-based human-machine interfaces. The benefit of using a systematically collected dataset is that it allows for an accurate and valid comparative study of various machine intelligence approaches for the human-robot interfacing. Without such a benchmarking approach, various factors, including the collection device and the condition, could affect the reported outcomes. We mainly focused on the second sub-database of the NinaPro database (DB2) [12], which has 17 hand movements. The hand gestures are depicted in Fig. 1.

A. Data Acquisition Process

The data consists of 40 intact subjects (28 males, 12 females; 34 right-handed, six left-handed; age 29.9 ± 3.9 years). The subjects were asked to hold the hand gesture for 5 seconds, followed by a rest of 3 seconds. This process is repeated six times for each hand movement, and the EMG signals were recorded from 12 Delsys Trigno electrodes. Due to delay in the subject's reaction time, the data at the end of the 5-second window may misrepresent the actual muscle movement. In order to counter this, the EMG signals are refined by relabelling the EMG signals. The electrodes were strategically placed around the forearm of the subjects. Eight electrodes were wrapped around the radio-humeral joints, two around the biceps and triceps, and further two around the flexor and extensor digitorum superficialis. The EMG signals were sampled at 2 kHz frequency with a baseline noise of fewer than 750mv RMS. These signals were filtered with Hampel Filter to remove 50 Hz powerline interference [12].

B. Data Preprocessing

We used minimal preprocessing of EMG signals to maintain the information content of the signal. The data is first normalized by using Z-score normalization with zero mean and unit standard deviation. The normalized data is then further rectified to transform the negative values to their absolute values. Signals are then windowed, and labels are assigned to each window. There is a trade-off between the length of the window for processing of the signals and predicting the intended gesture and the accuracy of the model. The longer the signal length, the more information is available to the model to make the prediction, but this would result in a prolonged delay in the system. The practical constraint about the agility of the system requires limiting the window length. Based on the literature, window of 300 ms [25] is considered to be acceptable for peripheral human machine intelligence, with 10 ms of stride to generate the training and test samples. No extra lowpass filtering is applied. As suggested in [12], for the first part of the work (before generalization) we used repetitions 1,3,4 and 6 for training purposes and repetitions 2 and 5 for testing.

III. Model Architecture

The model architecture has two stages: the LSTM stage and the CNN stage. In the LSTM stage, four LSTM layers are stacked together, each having 128 hidden parameters. Since the length of the input signal is of shape (600, 12), there are 600 LSTM units in each layer, as shown in Fig. 2. The last layer of LSTM outputs

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

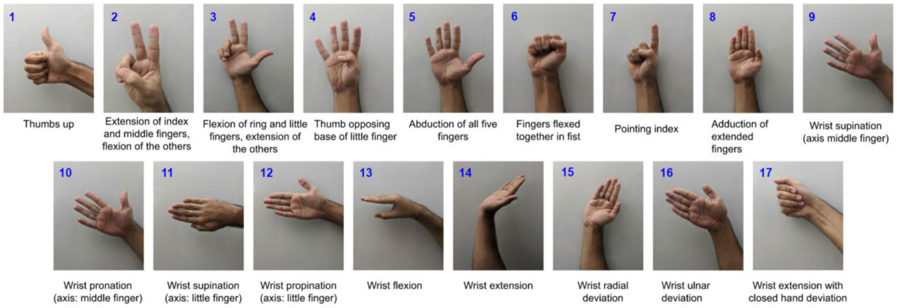
1
Thumbs up
2
Extension of index and middle fingers, flexion of the others
3
Flexion of ring and little fingers, extension of the others
4
Thumb opposing base of little finger
5
Abduction of all five fingers
6
Fingers flexed together in fist
7
Pointing index
8
Adduction of extended fingers
9
Wrist supination (axis middle finger)
10
Wrist pronation (axis: middle finger)
11
Wrist supination (axis: little finger)
12
Wrist propination (axis: little finger)
13
Wrist flexion
14
Wrist extension
15
Wrist radial deviation
16
Wrist ulnar deviation
17
Wrist extension with closed hand deviation

Fig. 1. Hand gestures to be classified.

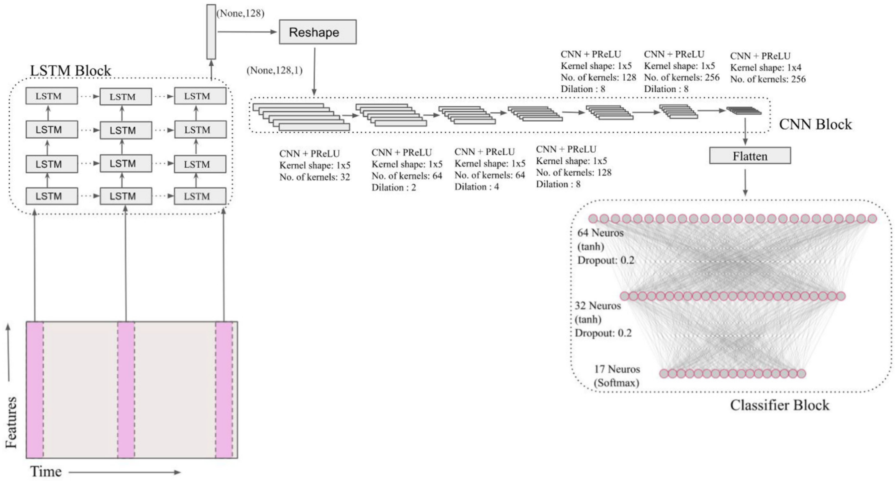

Fig. 2. Model Architecture.

a vector of shape (None, 128), which is then reshaped to (None, 128, 1) before feeding it to the CNN stage. There are 7 CNN blocks in the CNN stage. Every CNN block has three layers: Convolution Layer, Batch Normalization Layer, and PReLU layer. In this letter, we also added a dilation mechanism to the CNN part of the design. This is to enhance the memory of the CNN and allow for having heterogeneous access not only to short term temporal dependencies in the signal but also to long term dependencies. However, a dilation rate higher than eight is not used since a higher dilation rate will skip a high number of neurons and may cause loss of useful information. The last CNN block is followed by a classifier that consists of three fully connected layers, each having 64, 32, and 17 units, respectively. The first two layers have "$tanh$" activation function, and the last layer has "$softmax$" activation function, which produces the probabilities for each class. The classifier block is represented in Fig. 2. Using this approach, as explained in Section III, the average classification accuracy of 81.96% is achieved with averaged precision of 82.47% and sensitivity of 81.94%. The confusion matrix is shown in Fig 9. Also, the precision, recall, and F1 score are given in Table I.

We conduct a comparative study to evaluate the performance of the proposed hybrid method with a conventional CNN with a comparable number of trainable parameters of ≈1.4 M (for the Hybrid model we had ≈1.1 M parameters). The networks consists of two CNN blocks, each having one convolutional layer, one batch norm, one PReLU. There is a MaxPool layer between two CNN blocks with filter size of 2x2. Convolutional layer in the first block has 32 filters with the dimensions of 15x5; and the second CNN block has 64 filters each having size of 15x4.

TABLE I
RESULTS FOR USER-SPECIFIC GESTURE CLASSIFICATION

Methods	Accuracy	Precision	Recall	F1 Score
Hybrid	0.82	0.82	0.82	0.82
CNN	0.77	0.79	0.77	0.77
SVM	0.23	0.24	0.22	0.22

The 2D data was given to the CNN as the input by considering time and channels as the two dimensions. The CNN (as a conventional Deepnet method) significantly underperforms the proposed hybrid method by having an average accuracy of 77.30%, the precision of 78.80%, sensitivity of 77.38% over 17 gestures. A SVM model is also compared with the hybrid model, respectively achieving an average performance of 23.14%, 23.63%, 22.43%, and 22.33% in terms of accuracy, precision, recall, and F1-score. This shows the superior performance of the proposed hybrid approach.

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

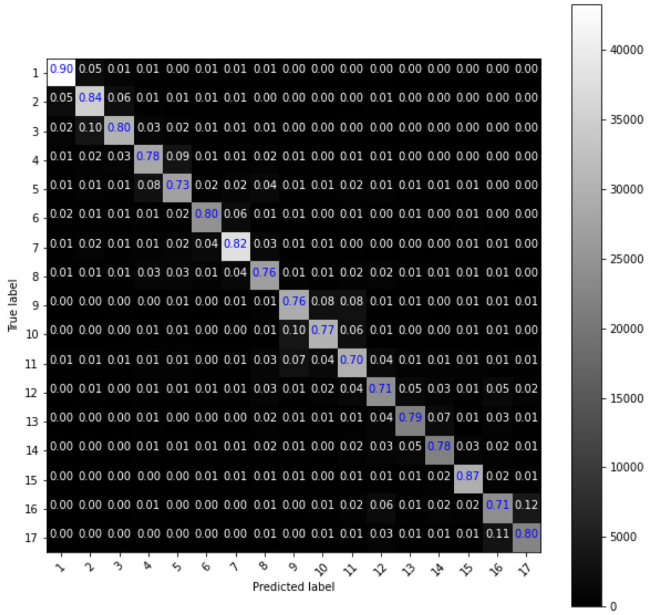

Fig. 3. Confusion matrix for 17 gestures.

IV. Generalization

Although there is existing research on the use of various deep learning algorithms, almost all existing approaches require a large number of repetition of each user as the models are trained to be user-specific, and the conventional goal was to detect a large number of classes for each user separately. This letter proposes to detect a lower number of classes but generalize it for all users so that machine intelligence is forced to learn the common underlying mechanism of the targeted gestures relaxing the need for calibration. If this goal is achieved, a major problem in the practical uses of neurorobots is addressed. Thus this letter, for the first time, aims to solve the generalized problem. We will later discuss the explainability of AI. In this work, we selected four gestures: wrist supination (axis: little finger), wrist pronation (axis: little finger), wrist flexion, and wrist extension.

A. Repetition-Based Generalization

First, we consider a compounded training dataset based on EMG data from all subjects for the selected gestures to create a sizable training and test data sets and training the machine intelligence in such a way that can generalize the repetition (in the next subsection we generalize based on subjects). The repetitions 1,3,4,6 from all subjects compounded were used for training the network, and the repetition 2 and 5 were used for testing the model accuracy. The architecture in Fig. 2 is then trained on the training data that constitutes EMG Data of from 40 subjects. The model achieved an accuracy of 79.33% on test data for the four selected gestures, while showing averaged specificity of 93.07% and averaged sensitivity of 80%. The confusion matrix for this task is given in Fig. 4. The performance of our proposed model is compared with CNN and SVM in Table II.

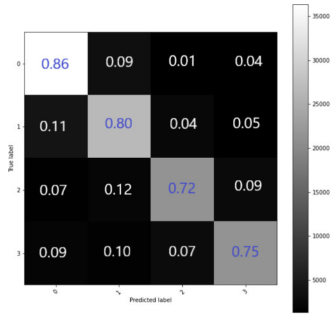

Fig. 4. Confusion matrix (Generalized Model) before optimization.

TABLE II
Results for Repetition-Based Generalization

Methods	Accuracy	Precision	Recall	F1 Score
Hybrid	0.79	0.78	0.80	0.79
CNN	0.70	0.71	0.70	0.70
SVM	0.47	0.48	0.44	0.44

This result show that the proposed hybrid machine intelligence algorithm is able to generalize the learned behavior across all subjects and detect the very fine common underlying neurophysiological behavior which represents the targeted gestures and relax the need for calibration to some considerable extent. This is for the first time that such a result is achieved. This result is extended in the next subsection using the second method of generalization.

For repetition-based generalization as can be interpreted from Table II, the average performance of the proposed hybrid method is as follows: precision of 78.25%, recall of 80%, and F1 score of 79%. The model achieved an accuracy of 79.3%. To compare with conventional methods we analyzed the performance of (a) an SVM model and (b) a CNN model. For the SVM we extracted 192 features from each sliding window of 300 ms with a step of 10 ms, including 48 temporal features and 144 spectral features. The temporal features include, four moments (mean, variance, skewness, and kurtosis) and the spectral features includes the same four moments for power spectrum density of the signal on the following frequency bands 0.5-12 Hz, 12-35 Hz, 35+Hz (for each processing window of 300 ms). Then, Principal Component Analysis (PCA) has been conducted to reduce the feature dimensionality to 18 features that represent 95% of the information from the original 192 features, given by the analysis of Proportion of Variation (PoV). For the CNN, we utilized the same model architecture as mentioned in the second paragraph of Section III. The average performance scores for SVM are as follows: accuracy of 46.95%, precision of 48.26%, recall of 44.14% and F1-score 44.29%. When compared with the results of the proposed hybrid method given above, it can be seen that the SVM results in a poor performance in terms of generalization showing the complexity of the problem. The CNN achieved the following average performance scores: accuracy of 70.26%, precision of 70.75%, recall of 70%, and F1-score of 69.5%. As can be seen the proposed hybrid method outperforms the

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

TABLE III
RESULTS FOR SUBJECT-BASED GENERALIZATION

Methods	Accuracy	Precision	Recall	F1 Score
Hybrid	0.77	0.77	0.77	0.76
CNN	0.71	0.71	0.70	0.70
SVM	0.32	0.29	0.29	0.27

conventional CNN when both are trained using the proposed repetition-based generalization methodology.

B. Subject-Based Generalization

In the previous subsection, we showed the performance when data from all subjects were combined in one pool to give the network a holistic view about the underlying common neural drive for the selected four gestures of all subjects. As a more rigorous type of generalization, named here as the subject-based generalization, we split the test and train subjects and completely isolate their repetitions to challenge the network towards zero calibration for unseen subjects. The total of 40 subjects was split into 26 of train and 14 of test subjects (maintaining a split similar to repetition-based generalization $\approx$ 3:1). No overlap exists between the train and test subsets. The model was able to achieve an accuracy of 77.17%. Average precision, recall and F1 score achieved are of 77%, 77%, and 76% respectively.

This result shows that thanks to the power of deep learning, instead of generating finely tuned models for detecting a high number of gestures for one subject, we can produce a generalizable model that can robustly detect a limited number of gestures but without having any observation/calibration-data from the new subject. This shed light on a new direction in human-machine interfacing, when generalizability (which allows for a smoother transfer to real-life application) is valued more than the number of classes (which cannot be realized in real-life). We compare the performance of our hybrid Deepnet technique with CNN and SVM. The results are added in the Table III. As can be seen in the table, the proposed approach outperforms CNN and the SVM. The Table III shows that basically SVM failed to solve this type of generalization, while the proposed hybrid approach secured a high precision, accuracy, recall and F1 score. It should be emphasized that main result here is to illustrate that generalizability can be achieved using deep learning techniques, which opens new doors for the future of calibration-free neurorobotics.

V. GRADCAM ANALYSIS

In this letter, to optimize the structure of the network and explain the performance, GradCAM algorithm is used. GradCAM is a method that visually explains how the model reached the decision for a particular class. It is often used in various computer vision applications to visualize the intermediate layers and how the neurons respond to certain inputs [24]. In this letter, we used the GradCAM method to 1) demystify the attention of the proposed model to various parts of the signal 2) detect the parts of the network which contribute the most to the classification (the focus of the model varies among CNN layers and is unique in different hand motions, revealing the unseen neurophysiological activity), and 3) use the knowledge to reduce the size of the network and the number of trainable parameters and to reduce the complexity of the deepnet. In the optimization step one CNN block which shows minimum contribution through GradCAM analysis was removed, reducing the size of the model and trainable parameters by 20% making the network shallower while having almost no effect on the performance. The spectrogram analysis is added only to better visualize the mentioned features and highlights the behavior of the network in the context of the GradCAM.

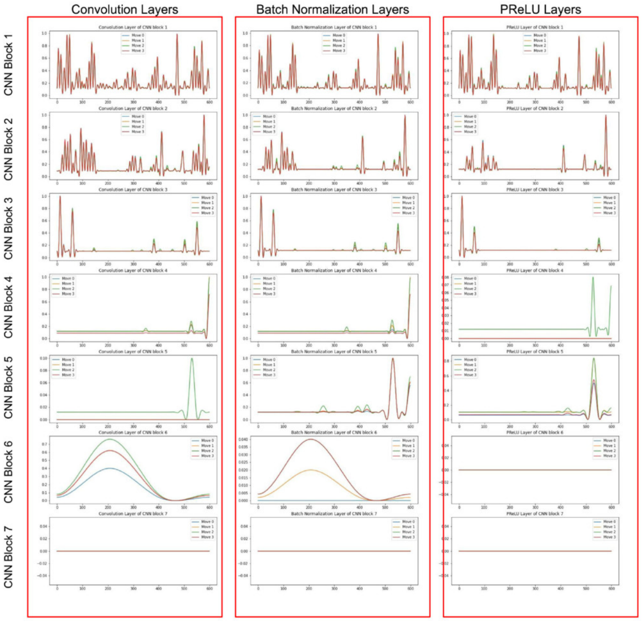

Fig. 5. GradCAM analysis performed on model with seven CNN blocks. Each row represent a CNN block and each column represent a different layer in the block. First layer represents convolution layer, second layer represents Batch Normalization, and third layer represents the PReLU layer.

Although usually only the last CNN layer is utilized in the literature about GradCAM, since the last layer has the most abstract and high-level information, we visualized all the layers to investigate the learning process. We utilized GradCAM for the generalized model and visualized all CNN, Batch Normalization and PReLU layers. The visualization results are shown in Fig. 5.

The GradCAM analysis in Fig. 5 shows seven rows corresponding to seven CNN blocks and three columns for three layers in each block. The first column is the Convolution layer, the second column is the Batch Normalization layer, and the last column is the PReLU layer. In the literature the rectified convolution layer that is considered since it is closest to the output for the GradCAM analysis. Thus, the most important plot of Fig. 5 is the last column and last row.

Through a simple visual inspection, it can be seen that the the last CNN block (last three layers) is not contributing much to the decision making. This can be observed since the four targeted gestures are not distinguishable based on the result in the last row. Thus it can be concluded that these layers are

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

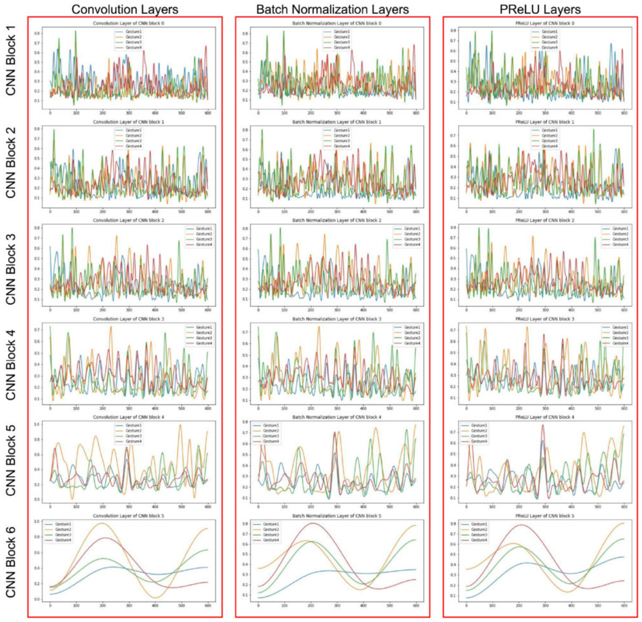

Fig. 6. GradCAM analysis performed on optimized model. Each row represent a CNN block and each column represent a different layer in the block. First layer represents convolution layer, second layer represents Batch Normalization, and third layer represents the PReLU layer.

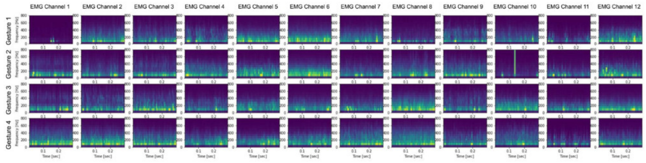

Fig. 7. Spectrogram of EMG Gestures. Each row represents a different gesture and each column represents spectrogram of a particular EMG channel. First to twelfth columns represents EMG channel 1 to 12 respectively.

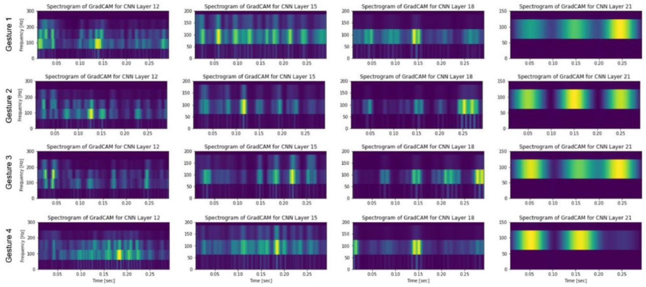

Fig. 8. Spectrogram of GradCAM signals. Each row represents a different gesture and each column represents spectrogram of a particular CNN layer. First to fourth columns represents CNN layer 12, 15, 18, and 21 respectively.

not adding value to the classification model of the intended geature while increasing the number of trainable parameters of the model and complexity of the model. In order to make the model less complex and reduce the number of trainable parameters, enhancing the practicality of the proposed solution, based on the above-mentioned analysis, we removed the last CNN block and trained the model again for generalization. The average accuracy and specificity and sensitivity of the model after optimization are 79.40%, 93.07%, 78.75%, respectively. The confusion matrix of the new model also can be seen in Fig. 9. Comparing the aforementioned results with that of the previous model, which had one more layer of convolution, it can be observed that the model, as predicted, has preserved the quality of performance in terms of accuracy, sensitivity, and specificity, highlighting the importance of the propose GradCAM analysis which led to optimizing the network. The new model, although securing a similar level of accuracy, has about 20% less trainable parameters and trained much faster as compared to the previous model, which is an imperative factor for real-life applications. Also, a smaller model generally requires less number of data for training. Thus, the GradCAM analysis resulted in designing a new hybrid approach with high accuracy and lower complexity.

The GradCAM analysis of each layer from the optimized model is shown in Fig. 6. When compared with Fig. 5, we observe that there are much abstract information and corresponding temporal activity in each layer as the model tried to predict a particular class. Also, the classes are more distinguishable in last CNN block, supporting the use of the optimized model. This can be seen by investigating the last row of Fig. 6, which is the most important part of this figure as it shows the last rectified convolution feature maps for GradCAM analysis, which contains most of the information about the discriminative power of the model. As can be seen in the last row of Fig. 8, the four classes are even visually separable, clarifying the strength of the proposed model to detect the corresponding gesture for a wide range of users generalizing the use of EMG-based interfacing.

VI. Spectrogram Analysis and Neural Code Using GradCAM Analysis

To further shed light on the performance of the proposed hybrid generalizable optimized network, we investigated the top 10 subjects (in terms of accuracy) to analyze the frequency spectrum of the EMG signals for each gesture. We used Hamming windows of size 32 with an overlapping of 28 (87.5% overlapping) to calculate the spectrogram of each gesture. The results are shown in Fig. 7.

As a result, Fig. 7 shows the spectrogram analysis of all the 12 channels for each of the four gestures. There are four rows and 12 columns. Each row represents a different gesture and each column represents one of the 12 channels of EMG inputs. This figure shows that most of the EMG signals have frequency components lying between 0-200 Hz range (consistent with the literature). However, as can be seen in the figure, there are no clear, distinguishable behavioral differences in the spectrogram of the EMG signals showing the complexity of the task. In the next step, we utilized the output of GradCAM analysis in the format of processed signals (as shown in Fig. 6). Here we calculate the corresponding spectrogram of the output activity of each layer of the network, and the results are given in Fig 8. Fig. 8 shows the spectrogram analysis of GradCAM signals (output of GradCAM analysis) of the last four convolution layers for all gestures. There are four rows, each representing a different gesture, and four

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

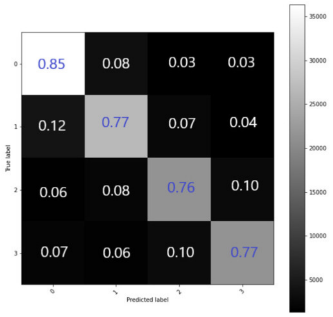

Fig. 9. Confusion matrix (Generalized Model) after optimization.

columns, each representing a different CNN layer. First column represent the spectrogram of GradCAM signals obtained for CNN layer 12. Similarly the second, third and fourth column represents the spectrogram of GradCAM signals obtained for CNN layer 15, 18, and 21 respectively. The results show that the network was able to process the spectrotemporal information and generate a distinguishable neural code for each gesture which can be even visually separated as can be seen in the last column of Fig.10 which explains the functionality of the neural network and the resulting assigned code by the proposed network to each layer. The aforementioned code indeed corresponds to the underlying neurophysiological activity, which is detected using the proposed approach allowing for generalizability of intention detection, minimizing the need for recalibration, which is an unmet need in the area of neurorobotics.

VII. Discussion and Limitations

In this letter, for the first time, we explore the generalizability of subject-wise hand gesture classification, and we propose to train a novel hybrid machine learning approach on the generalized problems in which the number of gesture classes is reduced, but the need for re-calibration for new and unseen users is dropped. This is done in contrast to the classical approaches, which solve a large number of gestures but under highly controlled conditions and only for one specific subject in one session.

In this work, we selected the dataset that includes 40 able-bodied subjects with different biomechanics to consider variability in terms of neurophysiology. However, it should be noted that this population does not reflect the biomechanics of people with the lack of a biological limb; since the biomechanics would be affected by the amputation surgery, and this can be a possible source of uncertainties. At the same time, as part of the amputation, the length of the muscle will be fixed by suturing them to the bones. It is worth noting that, the variation of the muscle length in able-bodied users is one of the major challenges as it results in a time variable volume conductor effect, which could significantly affect the signal linearity and stochastic behavior for able-bodied users. Analyzing the aforementioned opposing variables requires separate data collection and study. In this regard, to translate the results into practical applications, there is a need to conduct data collection from amputees and evaluate the performance of the proposed method on the corresponding dataset. This forms part of our future work. Besides training on data collected from amputees, in the future work, we will conduct research to discover more potential of the proposed hybrid approach and the training methodology by considering the higher number of gestures, different types (such as grasp and various mix of wrist and fingers motions), different combinations of gestures, and data collected from day-to-day variability to enhance the robustness and versatility of the generalized method.

VIII. Conclusion

In this letter, a hybrid LSTM-CNN model was proposed. The architecture was initially validated for detecting 17 classes of gestures in a user-specific manner (when trained over personalized data). The model secured an accuracy of 81.96%. As the next step, the model was generalized for relaxing the user-specificity of the algorithm, maximizing the practical uses for neurorobotics. Data of 4 gestures from all 40 subjects were combined into a single training data set. The generalized model achieved an accuracy of 77% for four gestures of unseen users. The GradCAM analysis was used to analyze the activity in each layer of the CNN block. Thus, the last CNN layer showed minimum contribution towards the decision-making. After removing the last CNN block to optimize the architecture using a shallower design, the model preserved the accuracy, specificity, and sensitivity. GradCAM helped in reducing the trainable parameters by 20%, thus bringing down the training time and memory consumption. The spectrogram plots for the GradCAM's output showed that the proposed network was able to transform the raw EMG activity into distinguishable neural codes accessing underlying neurophysiological activity based on which the network was able to generalize the intention classification problem for 40 users minimizing the need for calibration and addressing an unmet need in the area of neurorobotics.

References

[1] M. Stachaczyk, S. F. Atashzar, and D. Farina, "Adaptive spatial filtering of high-density EMG for reducing the influence of noise and artefacts in myoelectric control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 7, pp. 1511–1517, Jul. 2020.

[2] M. Stachaczyk, S. F. Atashzar, S. Dupan, I. Vujaklija, and D. Farina, "Toward universal neural interfaces for daily use: Decoding the neural drive to muscles generalises highly accurate finger task identification across humans," *IEEE Access*, vol. 8, pp. 149 025–149 035, 2020.

[3] S. F. Atashzar, M Shahbazi, and R. V. Patel, "Haptics-enabled Interactive Neurorehabilitation Mechatronics: Classification, Functionality, Challenges and Ongoing Research," *Mechatronics*, vol. 57, pp. 1–19, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957415818300448

[4] S. F. Atashzar, M. Shahbazi, M. Tavakoli, and R. V. Patel, "A computational-model-based study of supervised haptics-enabled therapist-in-the-loop training for upper-limb poststroke robotic rehabilitation," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 2, pp. 563–574, Apr. 2018.

[5] S. F. Atashzar *et al.*, "A grasp-based passivity signature for haptics-enabled human-robot interaction: Application to design of a new safety mechanism for robotic rehabilitation," *Int. J. Robot. Res.*, vol. 36, no. 5-7, pp. 778–799, 2017.

[6] S. F. Atashzar *et al.*, "A new passivity-based control technique for safe patient-robot interaction in haptics-enabled rehabilitation systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 4556–4561.

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.

[7] S. F. Atashzar, J. Carriere, and M. Tavakoli, “Review: How can Intelligent Robots and Smart Mechatronic Modules Facilitate Remote Assessment, Assistance, and Rehabilitation for Isolated Adults with Neuro-Musculoskeletal Conditions, Frontiers?” *Front. Robot. AI*, Frontiers, vol. 8, p. 48, 2021.

[8] R. Merletti *et al.*, “Technology and instrumentation for detection and conditioning of the surface electromyographic signal: State of the art,” *Clin. Biomechanics* (Bristol, Avon), vol. 24, pp. 122–34, 2008.

[9] W. Guo *et al.*, “An enhanced human-computer interface based on simultaneous semg and nirs for prostheses control,” in *Proc. IEEE Int. Conf. Inf. Automat.*, 2014, pp. 204–207.

[10] Y. Fan and Y. Yin, “Active and progressive exoskeleton rehabilitation using multisource information fusion from EMG and force-position EPP,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3314–3321, Dec. 2013.

[11] S. Tam, M. Boukadoum, A. Campeau-Lecours and B. Gosselin, “A fully embedded adaptive real-time hand gesture classifier leveraging hd-semg and deep learning,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 232–243, Apr. 2020.

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

[13] E. Rahimian S. ZabihiS. F. AtashzarA. AsifA. Mohammadi, “Surface EMG-based Hand Gesture Recognition via Hybrid and Dilated Deep Neural Network Architectures for Neurorobotic Prostheses,” *J. Med. Robot. Res.*, vol. 5, pp. 1–12, 2020, doi: 10.1142/S2424905X20410019.

[14] X. Zhou *et al.*, “Gesture recognition with EMG signals based on ensemble RNN,” *Guangxue Jingmi Gongcheng/Optics Precis. Eng.*, vol. 28, no. 2, pp. 424–442, Feb. 2020.

[15] M. Jabbari *et al.*, “EMG-based hand gesture classification with long short-term memory deep recurrent neural networks,” in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 3302–3305.

[16] T. Triwiyanto, I. P. A. Pawana, and M. H. Purnomo, “An improved performance of deep learning based on convolution neural network to classify the hand motion by evaluating hyper parameter,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 7, pp. 1678–1688, Jul. 2020.

[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, Oct. 2019.

[18] A. Samadani, “Gated recurrent neural networks for EMG-based hand gesture classification a comparative study,” in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 1–4.

[19] Y. Hu *et al.*, “A novel attention-based hybrid cnn-RNN architecture for semg-based gesture recognition,” *PLOS ONE*, vol. 13, no. 10, pp. 1–18, 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0206049

[20] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 02, pp. 107–116, 1998.

[21] W. Geng *et al.*, “Gesture recognition by instantaneous surface EMG images,” *Sci. Rep.*, vol. 6, 2016, Art. no. 36571.

[22] Z. Ding *et al.*, “semg-based gesture recognition with convolution neural networks,” *Sustainability*, vol. 10, no. 6, p. 1865, 2018.

[23] E. Rahimian *et al.*, “Semg-based hand gesture recognition via dilated convolutional neural networks,” in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2019, pp. 1–5.

[24] R. R. Selvaraju *et al.*, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[25] B. Hudgins *et al.*, “A new strategy for multifunction myoelectric control,” *IEEE Trans. Bio-medical Eng.*, vol. 40, no. 1, pp. 82–94, Jan. 1993.

Authorized licensed use limited to: New York University. Downloaded on May 29,2021 at 02:36:31 UTC from IEEE Xplore. Restrictions apply.