# X-MyoNET: Biometric Identification Using Deep Processing of Dynamic Surface Electromyography

Qin Hu, *Graduate Student Member, IEEE*, Alireza Sarmadi, *Graduate Student Member, IEEE*, Paras Gulati, Prashanth Krishnamurthy, *Member, IEEE*, Farshad Khorrami, *Senior Member, IEEE*, and S. Farokh Atashzar, *Senior Member, IEEE*

***Abstract*— This article investigates the potential of surface electromyography (sEMG) as a new biometric modality and proposes a deep neural network architecture as the backbone of a *gesture-independent* personal identification system (PIS). This article focuses on the real-world translation of such a model through systematic optimization, which finds the minimum number of gestures and sensors needed for training. Focusing on "dynamic sEMG," our proposed method can successfully identify 40 subjects with an average accuracy of 97%. This is achieved when gestures are the same in training, validation, and testing (the subjects need to repeat a particular gesture among a set of seven known gestures as a passcode). In a more complex scenario, when training gestures differ from those in validation and testing, our model can achieve an average accuracy of 90%, demonstrating that the proposed model can extract the unique patterns to identify a user regardless of gestures. Taking advantage of gradient-weighted class activation mapping (Grad-CAM), we explore the attention of the model on segments of the spectrotemporal space of the input signals. Grad-CAM not only sheds light on sEMG-based personal identification by decoding and visualizing the unique user-specific neurophysiological pattern but also generates a 2-D spectrotemporal mask used to reduce the model complexity significantly. As a result of the systematic optimization and Grad-CAM analysis, our proposed identification method needs only 4% of data for training, boosting practicality. This article also reveals the robustness of the proposed model for cross-day evaluation. Finally, the comparative study shows the superiority of our proposed model over several state-of-the-art algorithms.**

***Index Terms*— Biometric identification, explainable artificial intelligence (XAI), gesture-independent identification, gesture-sensor optimization, reliable cross-day identification, surface electromyography (sEMG).**

Manuscript received 19 December 2023; revised 27 February 2024; accepted 20 March 2024. Date of publication 5 April 2024; date of current version 19 April 2024. This work was supported in part by the U.S. National Science Foundation (NSF) under Grant 2229697 and Grant 2121391, and in part by the NYUAD Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute under Award CG010. The Associate Editor coordinating the review process was Dr. Chengyu Liu. *(Corresponding author: S. Farokh Atashzar.)*

Qin Hu, Alireza Sarmadi, Prashanth Krishnamurthy, and Farshad Khorrami are with the Department of Electrical and Computer Engineering, New York University (NYU), New York, NY 11201 USA.

Paras Gulati was with the Department of Electrical and Computer Engineering, New York University (NYU), New York, NY 11201 USA. He is now with JP Morgan Chase, Jersey City, NJ 07310 USA.

S. Farokh Atashzar is with the Department of Electrical and Computer Engineering and the Department of Mechanical and Aerospace Engineering, New York University (NYU), New York, NY USA, and also with NYU WIRELESS and the NYU Center for Urban Science and Progress (CUSP), Brooklyn, NY 11201 USA (e-mail: f.atashzar@nyu.edu).

Digital Object Identifier 10.1109/TIM.2024.3384571

## I. Introduction

THE rapid development of the Internet has contributed to the accelerated growth of several research fields, including medical technologies. An example is the Internet of Things (IoT) and the Internet of Medical Things (IoMT), allowing remote access to personal data for remote assessment and monitoring on telehealth platforms. However, this has imposed risks on personal information, such as medical and financial records [1], [2], [3], [4], [5]. In the last decade, substantial novel techniques have been designed and developed to mitigate the risks mentioned above for personal identification/verification purposes. The conventional methods such as personal identification number and password have been shown security deficient due to the possibility of information leakage, breaches, and counterfeits [6], [7], [8]. On a larger scale, there have been several reports on data breaches from credit agencies and governmental information systems, exposing the information of millions of employees and customers [9], [10], [11]. Biological-featured methods that extract the physical characteristics of human bodies, such as features of the face, fingerprint, and iris, are conventional approaches of personal identification system (PIS) to protect information privacy [12], [13], [14], [15]. However, these physiological methods are susceptible to hacking since technological advances allow for duplicating face models using 3-D printers, hacking fingerprints through latex gloves, and copying the corresponding biological features using artificial iris contact lenses [16], [17], [18], [19].

As a result, it is necessary to produce new means of biometrics that provide a higher level of personalization and reduce the risk of hacking. Biometrics that express *behavioral features* such as electromyography, the electrical manifestation of muscle contraction [20], [21], suggest a solid approach to achieve information security robustness by the corresponding uniqueness. This is because neurophysiological responses [such as those captured by surface electromyography (sEMG)] are unique to individual users and inherently complex in nature, making forgeries and falsifications exceedingly difficult [22], [23]. The advantages of sEMG as a new modality for personal identification are not limited to its uniqueness. One of the significant benefits of sEMG is that despite other physiological biomarkers, the sEMG passcodes can be tunable by changing to different gestures (studied as gesture-dependent

identification, where the proposed model trains, validates, and tests on the same gestures). Thus, this physiological security layer can be reset by choosing different gestures or muscles of a user. However, conventional biometrics are visible, such that they cannot be revoked or detached from users once compromised. Furthermore, the quality and usability of sEMG are not restricted to the identification environment and the physical states of a user, such as skin texture and amputation. sEMG is also invariant to the mental states of a user [24], [25], [26]. However, it should be noted that due to neurophysiological complexity and potential context-based variability, even for one individual, fundamental research is needed to generate techniques that can robustly and consistently detect the underlying sEMG "signature" as biomarkers. This is the focus of this article.

A few relatively recent works have been conducted in the literature regarding sEMG-based human identification, motivated by the unique characteristics of this biosignal to prevent personal information leakage, spoofing attacks, and identity theft. The conventional machine-learning (ML) approaches that are based on extracting temporal and spectral features from sEMG are the most commonly-used methods [27], [28], [29], [30]. The extracted features are then fed into conventional classifiers such as support vector machines (SVMs) and linear discriminant analysis (LDA) to identify individuals. One of the main limitations of the aforementioned studies is the simplicity of the extracted features and models. Thus, most of such efforts were conducted on small datasets, making generalized identification systems less achievable. In this regard, due to the variability, nonlinearity, and complexity of sEMG, it is imperative to test the capacity of this biosignal on a relatively larger number of individuals and gestures to detect the unique underlying features.

In recent decades, researchers have leveraged the powerful feature extraction capability of deep learning (DL) models to solve complex tasks. However, few of them exploited these models in human identification. In [31], the denoised sEMG signals were fed into a convolutional neural network (CNN) to minimize data preprocessing and let the model learn the underlying neurophysiological patterns on its own. Additionally, some researchers have recently converted raw sEMG signals into 2-D spectrograms, concurrently analyzing temporal and spectral muscle behaviors and potentially extracting high-dimensional information for generating biomarkers. This concept has been investigated in [16], where continuous wavelet transform was used in conjunction with a CNN architecture. Although the recent use of deep neural networks may suggest good performance, the existing works suffer from low diversification regarding subjects and hand gestures, raising concerns about generalization to a higher number of people and different gestures. Also, the existing recent DL research in personal identification cannot explain the attention of the neural network, raising concerns about the black-box modeling (e.g., biases in the dataset), which can be challenging for identification tasks and can pose a risk to system attacks. Moreover, the use of large spectrotemporal input spaces results in computationally inefficient models, challenging practicality in terms of the size of the training set and the implementation of small and portable identification hardware.

Motivated by the points mentioned above, this article proposes an identification method that uses an explainable CNN-based framework with the optimized input size derived based on the gradient-weighted class activation mapping (Grad-CAM) attention-based analysis of the system. Our identification method is compact, practical, interpretable, and robust. Gesture and sensor optimization finds the minimum number of gestures and sensors for training enhancing practicality. Two protocols are proposed to evaluate our method. In Protocol 1, the proposed method trains, validates, and tests on the optimal gestures and achieves 97% accuracy averaged across 40 subjects. When training on the optimal gestures but validating and testing on the remaining, nonoverlapping gestures (i.e., 33 gestures) in Protocol 2, the proposed method can achieve an average accuracy of about 90% across the same 40 subjects. As mentioned earlier, DL models have usually been deemed a black box because they only show the final predicting results but not the evidence (on the inputs) for making predictions. In this article, for the first time, we exploit explainable artificial intelligence (XAI) to interpret the proposed CNN model's attention and visualize each subject's extracted underlying neurophysiological patterns using Grad-CAM in sEMG-based personal identification. This explainability analysis also helps generate a 2-D spectrotemporal segmenting mask to further shrink the input space and reduce the model complexity. The multiday evaluation shows that our proposed method can identify the same subjects on two different days. We compare this article and the existing state-of-the-art efforts in sEMG-based biometric identification in the aspects of the number of participating subjects, involved gestures, placed sensors, the signal type and length, the model type, the proposed method, the sensor type, and the intraday and interday (if applicable) performance. The comparison is summarized in Table I. The six main contributions of this article are as follows.

*Contribution 1:* This article solves a challenging problem of gesture-independent identification, for the first time, in which common user-specific neurophysiological patterns across gestures are utilized by the model for identification. This departs from the current trend in the literature when users would be identified through performing one or a sequence of fixed gestures (i.e., gesture-dependent identification).

*Contribution 2:* For the first time, this work mainly focuses on the dynamic state of the sEMG activation for personal identification. The dynamic state includes imperative information regarding user-specific motor unit recruitment "patterns." Unlike the literature, which mainly processes the most stable phase of sEMG at the steady state that can be prone to contraction intensity, this work analyzes the rich motor unit recruitment information to solve a more challenging gesture-independent biometric identification problem for the first time.

*Contribution 3:* For the first time, this article systematically analyzes the optimal selection regarding the number and locations of sEMG sensors across diverse muscle groups and the corresponding effect on the system performance. These analyses will be needed for hardware implementation and identification accuracy boost.

TABLE I
COMPARISON BETWEEN THE PROPOSED MODEL WITH THE STATE-OF-THE-ART EFFORTS IN BIOMETRIC IDENTIFICATION

| Paper | # Subs | # Moves | # Channels | Signal Type | Sig Len | Model Type | Method | Sensor Type | Test Acc |
|---|---|---|---|---|---|---|---|---|---|
| [16] | 21 | 1 | 8 | steady state | 1.5 s | tree; CNN | DWT and CWT; gesture-dependent | sEMG | 99.2% (intraday) |
| [27] | 22 | 1 out of 8 | 64 | complete | 3 s | KNN | feature engineering; gesture-dependent | HD-sEMG | 99.8% (intraday); 54.03% (interday) |
| [30] | 20 | 1 out of 34 | 256 | transient removed | 4.5 s | SVM | feature engineering; gesture-dependent | HD-sEMG | 60% (interday) |
| [32] | 24 | 1 out of 16 | 8 | transient removed | 5 s | Mahalanobis distance | DFT; gesture-dependent | sEMG | 90.3% (intraday) |
| **ours** | **40 (intraday); 20 (interday)** | **any 1 out of 7 known; any 1 out of 33 unknown; 1 for interday** | **12 (intraday); 256 (interday)** | **complete** | **1.5 s** | **CNN** | **XAI; gesture-independent** | **sEMG (intraday); HD-sEMG (interday)** | **97.0% (known, intraday); 87.8% (unknown, intraday); 80% (interday)** |

Note: #: Number; Subs: Subjects; Sig Len: Signal Length; Acc: Accuracy; s: Second; ms: Millisecond; KNN: K-Nearest Neighbors; DWT/CWT: discrete/continuous wavelet transform; DFT: discrete Fourier transform.

*Contribution 4:* For the first time, this article conducts a holistic optimization to find the minimum number of gestures for training through clustering based on the synergistic similarities among sEMG signals from different gestures.

*Contribution 5:* This is the first article that implements optimal XAI (i.e., Grad-CAM analysis) to demystify an interpretable CNN model by processing sEMG for personal identification. The derived model's attention interprets and extracts the unique neural code in a visualizable and reportable manner projected on spectrograms of sEMG. Moreover, a particular spectrotemporal mask is proposed based on the attention of the network to reduce input space. As a result of gesture and sensor optimization and the spectrotemporal mask application, only 4% of training data are needed for identification, pushing our research toward developing portable personal identification security hardware.

*Contribution 6:* This article investigates the generalizability of the proposed method over two days and, at the same time, evaluates the performance on a single-day dataset. This approach would allow us to test the behavior of the model using two different platforms and under two different experimental settings, highlighting translation to practice and evaluating any unpredictable behavior under new situations. Such a systematic and comprehensive evaluation in the context of user identification is conducted for the first time in this article.

The rest of this article is written as follows. Section II introduces the data acquisition and preprocessing. Section III provides details on the methods. The results are presented in Section IV. Section V highlights the superiority of our proposed model over commonly-used classic and DL models. Lastly, concluding remarks are provided in Section VI.

## II. BIOMETRIC DATABASE

### A. Data Acquisition Process

Ninapro DB2, a publicly available open-source database, [33] is used in this article to evaluate the efficacy of the proposed methodology. The data collection was based on the Delsys Trigno system with 12 wireless electrodes, out of which eight channels are placed around the forearm near the radio-humeral joint, two are placed near the wrist on the extensor digitorum and flexor digitorum superficialis muscles, and two are placed on the biceps and triceps brachii muscles. Fig. 1 shows the electrode placements.

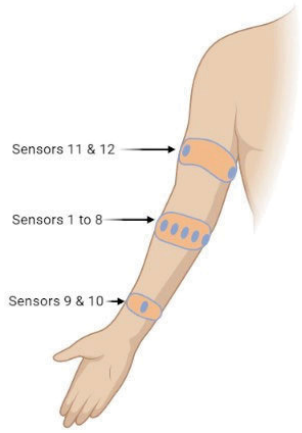


Fig. 1. Placement of myoelectric sensors [34].

The sEMG signals are recorded from 40 intact subjects (12 females and 28 males) having age 29.9 ± 3.9 years. DB2 is segmented into three exercises: exercises B, C, and D. Exercise B contains 17 gestures, among which eight are various isometric and isotonic hand configurations, and nine are wrist movements. Exercise C contains 23 gestures of grasping everyday objects and other functional movements. Exercise D contains nine force patterns. We combine exercises B (17 gestures) and C (23 gestures) for our research, taking into account a total of 40 gestures from 40 subjects. Subjects performed each hand movement six times, holding the gesture for 5 s followed by a three-second rest. sEMG signals are sampled at a frequency of 2000 Hz. The motion labels are further refined [33]. The Ninapro data were preprocessed using a notch filter at harmonics of 50 Hz to remove the power-line interference. This article uses preprocessed data. For multisession evaluation on our proposed method, we use the "Hyser" database described in detail in Section III-E.

### B. Data Preprocessing

Normalization is one of the data preprocessing techniques to maximize the performance and training stability of any ML approach by keeping the input features on a common scale without distorting the general distribution and ratios of the raw inputs. The $z$-score normalization [35] has been commonly used to mitigate model learning problems, such as inconsistent feature scales and vanishing gradients. In this

article, we propose a new normalization pipeline that is based on the $\mu$-law transformation followed by $z$-score normalization to boost the model performance.

The $\mu$-law transformation [36], [37] is a logarithmic and nonlinear transformation. This transformation increases the distinguishability among sensors and has been widely used in speech processing, where a voice (like sEMG) is also the convolutional summation of signals. The $\mu$-law transformation follows the mathematical design given by:

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

where $x_t$ denotes a single data value and $\mu = 2048$.

The strength of the signals collected by each of the 12 sensors varies according to the level of muscle contraction. The $z$-score normalization puts these signals on a common scale. The mean and standard deviation are found from the training data. Specifically let $i$ be the sensor index, $x_t^{(i)}$ be the single data point from sensor $i$, and $\mu_{\text{tr}}^{(i)}$ and $\sigma_{\text{tr}}^{(i)}$ be the mean and standard deviation of the sEMG signals from sensor $i$ of only the training data, respectively. The $z$-score normalization is given as

$$z_t^{(i)} = \frac{x_t^{(i)} - \mu_{\text{tr}}^{(i)}}{\sigma_{\text{tr}}^{(i)}}. \tag{2}$$

The duration of hand movement for different gestures and different subjects varies significantly over repetitions. These variations are often not considered in the literature. However, including imbalanced data (i.e., dissimilar amounts of input signals of hand movement from each subject) in training the proposed model may inject extra information that can influence and even inflate the overall model performance on sEMG-based personal identification. Thus, we only keep the first 1.5 s of data for each repetition that contains the reaction (between visual stimuli and movement start), transient (between movement start and maintenance), and steady-state (during movement maintenance) parts of hand movement, named dynamic sEMG. Previous work [38] defined the transient phase as the first second of each repetition according to the accelerometer signals. In this regard, more than 66% of the dynamic sEMG is from the transient phase. *Taking into account the dynamic state of contractions allows us to detect user-specific patterns during dynamic motor unit recruitment, which can be a distinguishing factor for the understudied problem.* Unlike the literature, which mainly focuses on processing the most stable steady-state sEMG (which would need an algorithm to locate and cut out the stationary part of the signal from the transient), our solution is more inclusive and does not rely on finding the point of transition. In addition, even though real-time implementation is not a major issue in sEMG-based biometric identification, we intentionally train our proposed method on dynamic sEMG, especially including the reaction and transient signals, to reduce the processing overheads that exist in the literature, according to Table I. As a result, all subjects are represented with the same length of dynamic signals in the dataset. Windowing is a method of data augmentation on sEMG data, which is important for a model to learn the covariant neurophysiological features

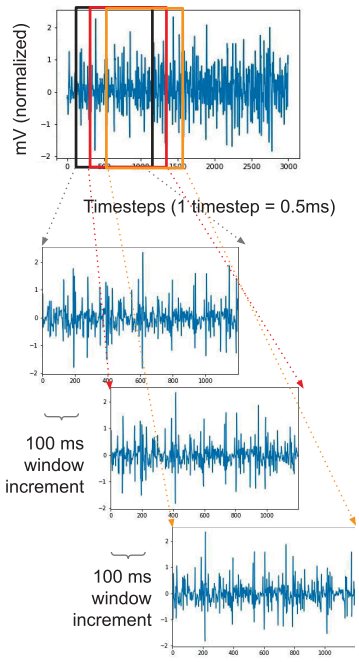


Fig. 2. Segmentation of sEMG signal into small windows.

regardless of phases of gesture performance. sEMG signals for each channel are segmented into windows (named "sliding windows"), each having a length of 600 ms and a stride of 100 ms. The windowing process is demonstrated in Fig. 2.

The 1-D sEMG signal from each electrode is converted into 2-D spectrogram images using short-time Fourier transform (STFT), which is applied with a window size of 500 ms and an overlap of 95%. The procedure of STFT is to segment each 600-ms sliding window of raw sEMG signals into small- and fixed-size windows (usually overlapping) and compute Fourier transform separately on each small window. This window size for STFT is a hyperparameter tuned based on the resulting spectrogram size and desired resolution in the time and frequency domains. Zero padding captures the information on the edge of each sliding window of raw sEMG signals. It is usually needed to obtain a reasonable-size spectrogram, especially when the window size for STFT is close to the sliding window size of raw sEMG signals. We also believe that a high-frequency resolution of a spectrogram will help generate a fine-grained spectrotemporal mask (discussed in detail in Section III-D3) that shows the contribution of different frequency bands when indicating the associated neurophysiological patterns with each subject. Hence, we choose 500 ms (1000 timestamps) to be the window size of STFT to reduce the need for zero padding but retain a high resolution of 2 Hz (2000 Hz sampling rate/1000 timestamps window size) in the frequency domain. We apply windowing before STFT to simulate the real-world implementation that an ideal human identification system should start identifying users with minimum data rather than waiting for the entire trial of sEMG. The raw spectrogram [shown in Fig. 3(a)] is clipped at 500 Hz [as can be seen in Fig. 3(b)] to discard potential noisy information above 500 Hz, serving as a low-pass filter. Thus, the final shape of the 2-D window (for each channel)

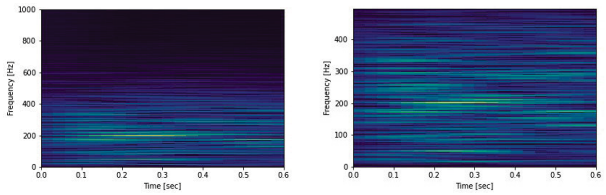


Fig. 3. (a) Spectrogram of a channel. (b) Spectrogram after clipping high frequencies.

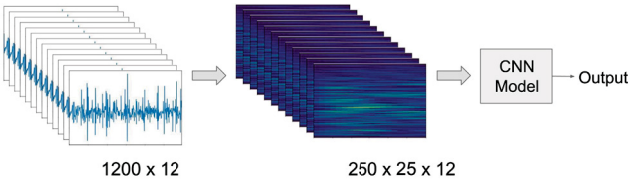


Fig. 4. High-level overview of the complete process.

after applying STFT and clipping can be represented by a 250 × 25 spectrogram, which is used later in this article as the input to the CNN model. The processing pipeline is summarized in Fig. 4.

## III. Method

A 1:*N* identification system extracts user-specific dynamic feature patterns to detect one user among a database of precollected templates of many users. Ninapro DB2 allows us to explore the possibility of sEMG as a biomarker without the restriction of extracting neurophysiological patterns from a limited number of gestures and subjects. In this regard, unlike the literature (see Section I) that identifies a user through one or a sequence of fixed gestures (i.e., gesture-dependent personal identification), this article conducts "gesture-independent" personal identification by evaluating our proposed method using two protocols. The goal of Protocol 1 is to train the proposed model using the same set of gestures in the training, validation, and testing phases, whereas, in Protocol 2, the model is further challenged by differing the training gestures from the validation and testing gestures, aiming for generalizability and forcing the model to learn common gesture-independent user-specific neurophysiological patterns for identification. This article aims to design a compact, optimized, explainable, day-to-day reliable, robust identification system. The methodology details about model architecture, gesture and sensor optimization, explainability-based optimization, and multisession evaluation can be found in Sections III-A–III-E.

### A. Initial Model Architecture

The proposed model takes commonly used spectrograms transformed using STFT as inputs for training, validation, and testing, processing the spectrotemporal dynamics in sEMG for biometric identification. The 2-D spectrograms (the third dimension corresponds to the sensors) generated in Section II-B are processed using the proposed neural network to detect the corresponding subject through a classification scheme. For this purpose, we exploit the power of neural networks for classification. Our model consists of two

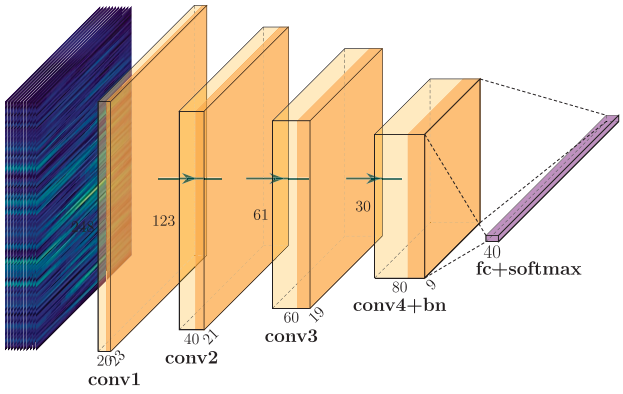


Fig. 5. Overall architecture of the model.

TABLE II
Model Architecture

| Layer Name | Layer Type | # of Channels | Filter Size | Stride | Activation |
|---|---|---|---|---|---|
| conv1 | Conv2d | 20 | 3 × 3 | 1 × 1 | ReLU |
| conv2 | Conv2d | 40 | 3 × 3 | 2 × 1 | ReLU |
| conv3 | Conv2d | 60 | 3 × 3 | 2 × 1 | ReLU |
| conv4 | Conv2d | 80 | 3 × 3 | 2 × 2 | ReLU |
| bn | BatchNorm2d | 80 | - | - | ReLU |
| fc | FC | 21600 | - | - | - |

modules: the autonomous feature extractor and the classifier. We choose a CNN as they have been shown to be powerful for feature extraction [39], [40]. Performing weight sharing through sliding kernels in a CNN results in a smaller number of trainable parameters. Also, by sliding a kernel in 2-D space, the CNN can detect the feature patterns appearing anywhere in a spectrogram. This lightweight but robust model structure is well-suited to our design of a compact identification system.

The proposed model contains four CNN blocks, each having a CNN layer (feature extractor) and a rectified linear unit (ReLU) activation function, followed by a fully connected (FC) classifier (Fig. 5). In order to improve the model convergence during training, a 2-D batch normalization is used in the last CNN block. The summary of the model is reported in Table II.

In all the experiments, the models are trained for a maximum of 500 epochs with a batch size of 32. Adam optimizer is used with a learning rate of 0.0001, which is reduced by a factor of 0.1 after the first 100 epochs.

### B. Gesture and Sensor Optimization

Relying on a large number of training gestures and sensors is not practical for personal identification in the real world. To enhance the practicality and usability of the proposed approach, we investigate how to reduce the number of: 1) training gestures and 2) sensors to find the optimal selection while obtaining similar performance to the larger number of gestures and sensors.

The intuition of gesture-based optimization of the input space comes from synergistic similarities among sEMG signals of different gestures that can be clustered into low-dimensional finite groups [41], [42], from each of which one representative gesture can be selected as a training gesture.

The sensor-based optimization of the input space is based on individual sensor performance ranking. It should be noted that we conduct gesture and sensor selection in a sequential manner, meaning that for the gesture-based optimization, sEMG signals from all sensors are considered, while for the sensor-based optimization, the optimal gestures from the previous step are considered.

*1) Gesture-Based Input Space Optimization:* For gesture-based optimization, we extract various features from both time and frequency domains to capture distinguishing spectrotemporal patterns. The temporal features are mean absolute value, variance, mean square root, root mean square, log detector, waveform length, difference absolute standard deviation value, zero crossing, skewness, and kurtosis [29]. For the spectral features, we considered conventional neural frequency bands of delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-35 Hz), and gamma (>35 Hz) [43], [44]. For each mentioned frequency band, mean power density is considered to be the spectral feature. We extract these features from each sensor on the sEMG signals after $\mu$-law transformation and averaged over subjects and repetitions. Thus, the preclustering data dimensions in the time domain are $40 \times 120$ (where 40 corresponds to the number of gestures and 120 corresponds to the ten temporal features calculated for every 12 sensors). Also, in the frequency domain, the dimension is $40 \times 60$, where 40 represents the number of gestures and 60 corresponds to the five spectral features calculated for all 12 sensors.

GMM [45], [46], [47] is used to cluster the gestures based on the extracted temporal and spectral features (180 for gesture clustering). In this article, GMM is initialized using $K$-means to increase the convergence speed and reduce some computational burden. Compared to $K$-means, which gives each data point (e.g., a gesture) a hard assignment to a particular group, GMM is a soft clustering approach that gives a probability to each data point belonging to a Gaussian component. Furthermore, the GMM parameters (mixture weights, means, and variances) are iteratively updated through the expectation–maximization (EM) algorithm to find the maximum likelihood of a GMM best capturing the distribution of the gesture representations. A GMM represents the distribution of gestures in the form of

$$p(f_\theta(x)|\lambda) = \sum_{k=1}^{m} \omega_k \cdot g(f_\theta(x)|\mu_k, \Sigma_k) \tag{3}$$

where $m$ is the number of the user-defined clusters, $f_\theta(x)$ is the representation of the gesture $x$, and $\theta$ is the parameter of the representation. $\lambda$ is the vector of GMM parameters including mixture weights ($\omega_k$), mean vectors ($\mu_k$), and covariance matrices ($\Sigma_k$). Gaussian densities ($g(f_\theta(x)|\mu_k, \Sigma_k)$) are given by

$$g(f_\theta(x)|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} e^{-\frac{1}{2}(f_\theta(x)-\mu_k)^\intercal \Sigma_k^{-1}(f_\theta(x)-\mu_k)} \tag{4}$$

where $D$ is the dimension of the representation.

Before implementing GMM, we first apply principal component analysis (PCA) [48] to reduce the $40 \times 180$ feature space (containing 40 gestures with temporal and spectral features) to $40 \times 15$ for better clustering results. GMM clustering requires the number of clusters $m$ as input. We implement widely adopted Bayesian information criterion (BIC) (see the following equation) to derive the optimal number of clusters as the minimal number of training gestures [49]:

$$\text{BIC score} = \log p(f_\theta(x)|\lambda) - \alpha \frac{1}{2}\beta \log N \tag{5}$$

where $\alpha$ is a penalty weight, $\beta$ is the number of parameters in a GMM model, and $N$ denotes the number of gestures.

A GMM becomes more complex when the number of Gaussian components increases, potentially resulting in an overfitting problem. As a BIC score is penalized by the model complexity (the number of components) in a GMM, we choose the number of clusters as seven that has the lowest BIC score to avoid overfitting problems. The GMM result shows that seven Gaussian components can optimally and sufficiently capture the distribution of the $40 \times 15$ feature space. This article selects one gesture with the highest log-likelihood from each assigned cluster as the representative training gesture of that cluster. As a result, we select seven optimal gestures (named "Optimal Gestures" in the rest of this article) to reduce the input space.

*2) Sensor-Based Input Space Optimization:* Sensor optimization employs performance ranking to find the optimal sensors. In this approach, we feed sEMG signals from Optimal Gestures and only one sensor at a time into the proposed model. The best-performing sensors are added one by one into the training set according to the individual performance in descending order for comparison. According to the one standard error rule, performance ranking returns five as the minimal number of sensors (named "Top Sensors" in the rest of this article) needed for training, securing almost similar performance with smaller input size. However, Top Sensors were sparsely positioned on all three muscle groups used in the dataset. To further enhance the practicality of our proposed identification method, we select five sensors (named "Optimal Sensors" in the rest of this article) positioned only on two muscle groups based on sensor performance ranking.

### C. Model Validation Protocols

This article evaluates the performance of the proposed model by two protocols based on the results of gesture and sensor optimization. In both protocols, sEMG signals of corresponding gestures are pooled together in training, validation, and testing, tackling gesture-independent personal identification tasks. The protocols are described in the following and can be visualized in Fig. 6.

*1) Protocol 1:* For this protocol, our model is trained, validated, and tested on the seven Optimal Gestures to identify 40 subjects. The training, validation, and testing are based on repetitions (2, 4, 6), (1), and (3, 5) of these gestures, respectively. A user can be identified by performing any one gesture of their choice from the seven Optimal Gesture Set, showing the user-friendliness of our proposed method.

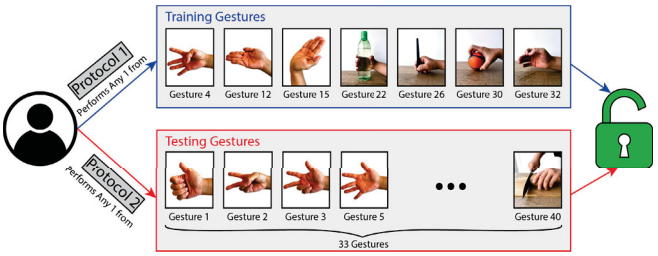


Fig. 6. Model validation protocols. In Protocol 1, subjects will be identified by performing any one gesture of their choice from the Optimal Gestures or training gestures (in the blue rectangle). In Protocol 2, they will be identified by performing any one gesture from the testing gestures (in the red rectangle), which are unknown to the proposed model during training.

*2) Protocol 2:* In this protocol, which is more challenging from a ML perspective, the training gestures (i.e., seven Optimal Gestures) differ entirely from validation and testing gestures (i.e., the remaining 33 gestures in the database). The training is based on all six repetitions of Optimal Gestures, while the validation and testing are based on repetitions (2, 4, 6) and (1, 3, 5) of the remaining gestures, respectively. We hypothesize that common underlying neurophysiological patterns can be found across gestures for a user. Learning and extracting these patterns can prevent overfitting to specific user inputs (e.g., sEMG collected at a certain hand angle or muscle contraction level), counteracting sEMG variation because muscle contraction can be different from the same user when performing the same gestures at different times. Moreover, Protocol 2 further enhances user-friendliness and practicality by allowing users to perform any gesture from the 33 gestures not used in the training set, freeing users from learning a standardized way of doing a particular gesture(s).

### *D. Grad-CAM Analysis: Explainability-Based Optimization*

Grad-CAM enhances the transparency and explainability of a black-box CNN-based network through the gradients of any given class flowing into the last convolutional layer of the network, producing a heatmap that highlights the network attention on the input [50]. As a part of XAI, Grad-CAM is often used to reveal the attention of machine intelligence, extract the underlying information invisible to the naked eye, and optimize the size of the dataset and model architecture. This article uses Grad-CAM to: 1) help visualize the attention of the network on the average subjectwise spectrograms and the corresponding localization maps in parallel; 2) extract the identification code from the overlay of the averaged spectrogram and attention heatmap for each user; and 3) extensively reduce the input spaces and the number of the trainable model parameters, optimizing the size of the proposed network.

*1) Subjectwise Attention Heatmap Generation:* In this article, we concatenate Optimal Sensors horizontally to preserve the critical channelwise localization information and show the model's attention on different sensors. The Grad-CAM analysis is conducted on the best-performing model trained on Optimal Gestures and concatenated Optimal Sensors in the validation set to demystify the model decision. The horizontal concatenation broadens the input size of each channel (the third axis of inputs) from $250 \times 25$ to $250 \times 165$ with a zero padding of $250 \times 10$ in between sensors to generate distance between each sensor information. We do not want the proposed model to treat transitions as part of the signal patterns; hence, we have introduced gaps using zero padding. To achieve better model performance, we slightly modify the proposed model architecture by setting the stride as $2 \times 2$ in all convolutional layers. The gradients from the last convolutional layer are extracted and resized from $30 \times 19$ to align with the input size to form the attention heatmap. Each heatmap indicates the model attention on each sample spectrogram. The average spectrograms and corresponding heatmaps by subjects are presented in parallel as the results of the attention analysis.

*2) Identification Code Extraction by Subject:* By visual analysis of Grad-CAM, the model attention varies in sensors and frequency ranges for different subjects, later defined as frequency bins. The attention indicates the distinguishability and uniqueness of the spectrotemporal neurophysiological characteristics of each subject. These underlying sEMG features can be translated into identification codes that have the purposes of: 1) investigating the distinguishing neurophysiological patterns associated with each individual, further qualifying sEMG as a biomarker and 2) providing intuition and knowledge for the spectrotemporal mask generated by an automatic algorithm based on the attention heat of our proposed model, further reducing the model complexity by shrinking the input space. The identification codes from the same subject are highly consistent across repetitions.

To extract the unique identification code for each individual, we segment the average spectrogram of each subject ranging from 0 to 500 Hz into 25 frequency bins, each containing 20-Hz spectrotemporal information across sensors. Thus, an identification code is an array of five scalar values, named as identification scalar, each falling into a range between 1 and 25. An identification scalar is calculated through the equation given by

$$I(s) = \left\lceil \frac{\mathrm{GC}_{\mathrm{freq}}}{10} \right\rceil \tag{6}$$

where $\mathrm{GC}_{\mathrm{freq}}$ is the coordinate of a gravity center on the frequency axis of a heatmap and $s = 1, \ldots, 5$ is the index of the five Optimal Sensors. The function center_of_mass [51] is used to obtain the gravity center coordinates of the subjectwise Grad-CAM heatmaps of the sensors for each subject. We use the blurring and thresholding method to ensure the true gravity center at each sensor is precisely defined, reducing the noise. As the next step, we convert the heatmap to binary images to accentuate the hottest areas. It should be noted that this approach also avoids the gravity center shifts. In Fig. 7, before applying blurring and thresholding, the gravity center of the given sensor is found between 151 and 160 on the frequency axis, translated into identification code 16. However, after applying blurring and thresholding, the gravity center of the same sensor is found between 171 and 180 on the same frequency axis, translated into identification code 18. Thus, we observe a shift of two in identification code when the blurring and thresholding method is not applied.

However, this approach may fail to detect any hot zone when the heatmap is highly dispersed. The sensorwise segmentation

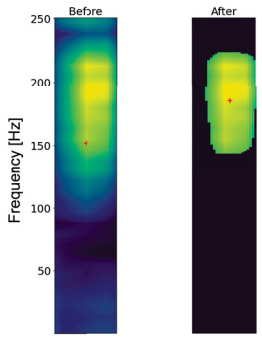


Fig. 7. Heatmap of a sensor before and after blurring and thresholding, subject 30 on sensor 8.

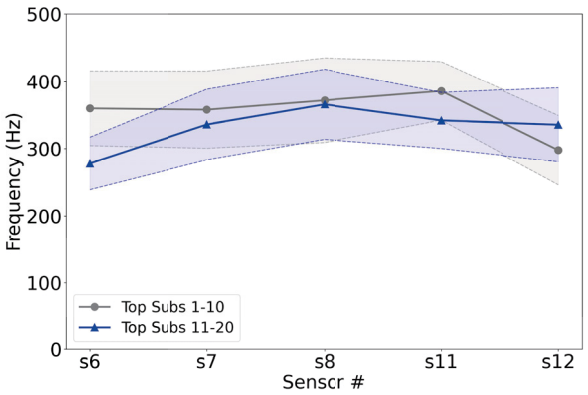


Fig. 8. Means and standard deviations of identification codes of top-performing subjects. Solid lines indicate the means. Dashed lines and filled areas indicate the distances of one standard deviation away from the means. The legend "Top Subs 1-10" means the top-1-to-10-performing subjects. The meaning of the other legend follows the same pattern.

of the attention heatmap is used as an alternative approach that further divides the heatmap corresponding to the sensor into five equal 100-Hz segments. Each segment returns an average heat, indicating the strength of the model's attention on that segment. By using the same function, five gravity center candidates can be found at each sensor. The gravity center of the segment that has the highest average heat is considered as the final candidate for the gravity center at the sensor. Thus, the gravity centers, respectively, represent the most concentrated attention spots for the best sensors on each heatmap. The resulting five centers form the identification code of a particular subject, which shows the specific attention of the network on different frequencies and sensors for identifying each subject.

To evaluate the relation between the model attention and the model performance (see Fig. 8), we calculate the means and standard deviations of the identification codes of the top 1–10 and top 11–20 performing subjects and convert the analysis results back to frequencies in hertz. The analysis results show that the proposed model pays attention to median-to-high frequencies, especially the higher gamma band of >80 Hz for best-performing subjects. This observation matches the observation on the spectrotemporal mask and serves as an examination of the automatic algorithm that generates the mask, more than 50% of which includes these frequencies.

*3) Attention-Based Spectrotemporal Mask Generation and Model's Size Optimization:* Based on the previously mentioned gesture-based and sensor-based optimizations, we enhance the practicality by minimizing the number of gestures and sensors used in training. The smaller input space and less trainable parameters can further refine the proposed identification system by reducing the data storage, speeding up the training process, and increasing the practicality.

In this section, we propose an optimizing attention-based spectrotemporal mask that abandons the trivial areas, which play the minimum role in classification from the input space. We hypothesize that retraining the model only on the most informative segments of spectrograms can result in similar performance while significantly reducing the model size. The median Grad-CAM heatmap of the top-10-performing samples of the spectrograms from the validation set is utilized to generate the most significant attention-based segment of spectrotemporal information across subjects and gestures. The median heatmap rather than the average heatmap is employed for mask generation as the data may not be normally distributed. Based on the results achieved on the model attention summarized from the previous radius distribution analysis, the optimizing spectrotemporal mask is systematically calculated using sensor segmentation. We segment each sensor into multiple fine pieces and select the top 60% segments that have the highest average heat. The outcome consists of both low- and high-frequency areas at each sensor (Fig. 9). It is expected that our model pays attention to both low- and high-frequency areas on the average spectrogram because our input signals include low-frequency contraction at transient phase and high-frequency contraction at plateau phase.

After applying the mask (calculated based on the average attention map of best-performing subjects) on each spectrogram for all subjects, the segments for each sensor are concatenated vertically, making one transformed spectrogram for each sensor. The resulting five transformed spectrograms are horizontally concatenated with zero paddings in between, forming the small input space (see Fig. 10) for the model. The model is retrained on the reduced dataset.

### *E. Evaluation on Multisession sEMG*

The multisession evaluation of our method is imperative for showing the suitability of sEMG as a biomarker. sEMG recordings from the same subject on varying days could be different due to the possible variation in neurophysiology, artifacts caused by stochastic noises, and electrode misplacement and displacement during doffing and donning [52]. We evaluate the robustness of the proposed model on a two-session dataset collected using high-density surface electromyography (HD-sEMG) from 20 subjects.

The conducted multisession evaluation is based on a publicly available HD-sEMG dataset ("Hyser") that includes 16 different degree-of-freedom finger and wrist gestures collected from two different days with a cross-day interval of 3–25 days [53]. The dataset was collected from 20 intact subjects, who maintained each gesture for 4 s for two repetitions each day. The HD-sEMG signals were acquired using the Quattrocento system (OT Bioelettronica, Turin, Italy) through four 8 × 8 electrode grids (a total of 256 sensors) with a sampling rate of 2048 Hz. On each forearm side (extensor

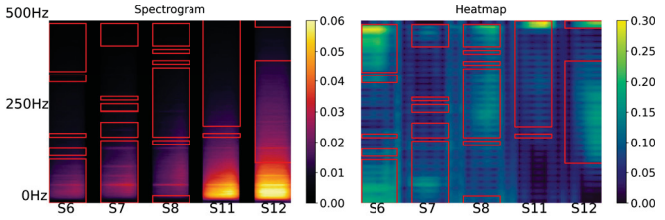


Fig. 9. Spectrogram segmentation and mask generation.

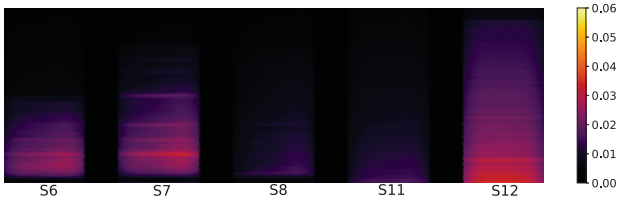


Fig. 10. Example: mask application result.

or flexor), two 8 × 8 electrode grids construct one 16 × 8 electrode grid. The Hyser data were first preprocessed using an eighth-order Butterworth bandpass filter between 10 and 500 Hz, followed by a notch filter at harmonics of 50 up to 400 Hz to remove the power-line interference. This article uses preprocessed data.

The number of sensors in Ninapro (with 12 sensors) is different from that of Hyser (with 256 sensors). To keep the model structure consistent regarding the number of input channels (sensors), we conduct an additional data preprocessing step on the high-density data. Thus, we apply a 2-D average pooling with a kernel and stride size of 4 × 4 to each HD-sEMG grid. The pooling outputs are flattened and concatenated to obtain 16 input channels of highly condensed and highly representative information. In addition, we reduce the window stride to ten timestamps to mitigate the overfitting issue caused by the 20 subjects difference between the two databases. We train the model for a single gesture on Day 1, validate the same gesture from the second repetition of Day 2, and test on the same gesture from the first repetition of Day 2. We also investigate which gestures (out of 16) can secure a reliable cross-day performance.

## IV. Results

This article uses accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve, and area under the curve (AUC) score averaged across subjects to evaluate the performance of the proposed model. These metrics are commonly used to comprehensively evaluate biometric identification (see examples in [54] and [55]). We implement the majority-voting strategy for the metric calculation to optimize the performance and practicality of our identification system. In this regard, our model predicts a subject for each repetition of gesture performance based on the predicted majority of its ten windows.

### A. Gesture and Sensor Optimization

In gesture-based optimization, we investigate five, six, and seven gestures and evaluate the corresponding model performance given all 12 sensors because of the similar BIC scores. According to the one standard error rule, we choose seven Optimal Gestures of 4, 12, 15, 22, 26, 30, and 32 based on the mixed-domain (temporal and spectral) clustering. The gesture numbers correspond to the thumb opposing base of the little finger, wrist pronation (rotation axis through the little finger), radial wrist deviation, medium wrap, writing tripod grasp, tripod, and tip pinch grasp.

In sensor-based optimization, we derive Top Sensors based on the ranking of individual sensor performance and then combine the most informative sensors. An identification system is easier to use when a user is required to attach sensors to fewer locations on the arm, potentially attracting more users. In order to further enhance the practicality of our identification system, our analysis results in sensor IDs 6, 7, 8, 11, and 12 to be the five Optimal Sensors spreading among two muscle groups (extensor–flexor group and biceps and triceps group), achieving the same accuracy as the five Top Sensors [see Fig. 11(b)].

*Remark 1:* It should be highlighted that our proposed method can identify 40 subjects after training only on 7% of data from Ninapro DB2 beneficial from gesture and sensor optimization.

### B. Results of Validation Protocols

For Protocol 1, in which training, validation, and testing are based on the same seven Optimal Gestures, our proposed model achieves 96.96% accuracy, 97.17% precision, 96.96% recall, 96.95% F1 score, and 0.998 AUC averaged across 40 subjects. Even though our proposed model is trained on Optimal Gestures but validated and tested on the 33 nonoverlapping gestures with a train-validate-test split of 18/41/41 in Protocol 2, it still achieves 87.75% accuracy, 88.21% precision, 87.75% recall, 87.78% F1 score, and 0.991 AUC across same subjects. By solving such a challenging task in Protocol 2, this article shows the power of the proposed model in extracting common underlying user-specific neurophysiological patterns regardless of gestures, indicated by the less than 10% performance reduction compared with the results of Protocol 1.

### C. Grad-CAM

Fig. 12 shows the spectrograms and attention heatmaps (generated by Grad-CAM) of two subjects (i.e., #19 and #28) with top performance. We can conclude that the proposed model makes decisions based on distinct frequency bins among sensors, which reveal the underlying spectrotemporal patterns that can be exploited to identify subjects and optimize the proposed model. As explained, the unique features of the attention heatmaps are translated into identification codes, each consisting of five frequency bin numbers ranging from 1 to 25. For example, the algorithm generates the identification codes 14-17-21-7-2 for Subject #19 and 25-25-25-11-4 for Subject #28.

We utilize Grad-CAM to further reduce the size of the network by optimizing the input space. The results show that the application of the spectrotemporal mask reduces the individual input size from 250 × 165 to 150 × 165. Hence, the proposed approach allows for dropping 40% of

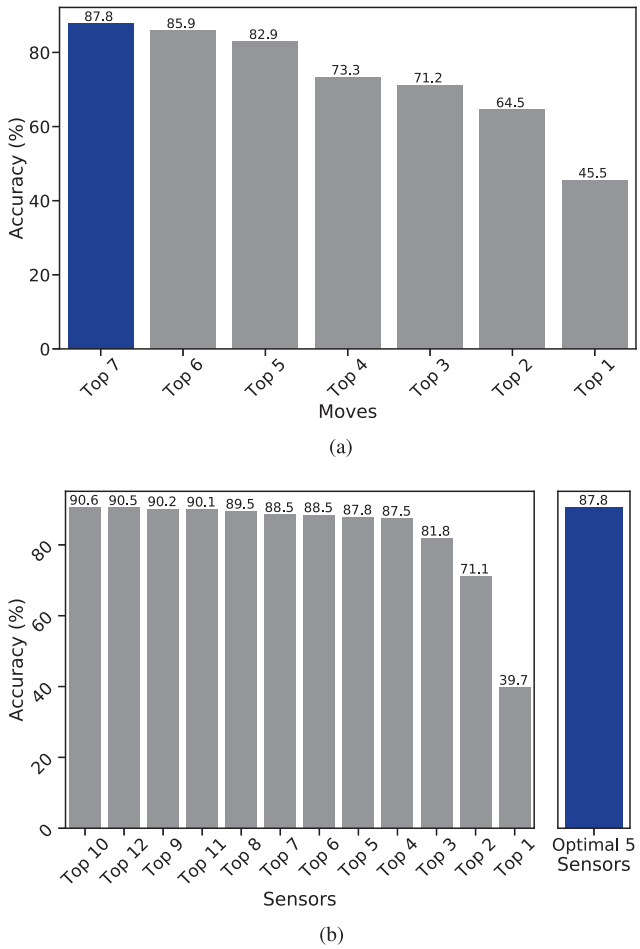


Fig. 11. (a) Performance ranking of top gestures is based on Optimal Sensors; the blue bar shows the accuracy of Optimal Gestures. (b) Performance ranking of Top Sensors is based on Optimal Gestures; the blue bar shows the accuracy of Optimal Sensors.

the trainable parameters of the network, from 938k to 563k parameters, resulting in a much less complex network. Also, it results in over 20% reduction of time needed for training (this number may vary on different machines). This approach achieves these reductions while accuracy, precision, recall, and F1 score are decreased by about 5%, and AUC by 0.005 compared with model performance in the optimization section (see Section IV-A). Fig. 13 shows the microaverage ROC curves across subjects before and after applying the spectrotemporal mask. The model performance after applying the mask shows the efficacy of the proposed attention-based data masking optimization technique proposed in this article.

### D. Evaluation on Multisession sEMG

We conduct a comprehensive analysis on each of the 16 gestures that have two-day data. We observe three gestures: 1) middle finger extension; 2) hand close; and 3) hand open (see Fig. 14) that show high reliability in distinguishing all 20 subjects with 80% average accuracy, 71.39% average precision, 80% average recall, 74.17% average F1 score, and 0.946 average AUC (see Fig. 15). These results prove that our proposed model trained on Day 1 can still identify subjects on Day 2 by robustly capturing the neurophysiological signature associated with each subject.

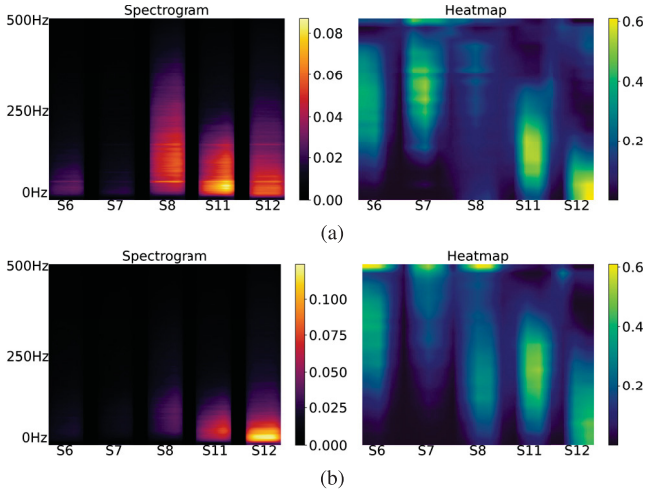


Fig. 12. Average spectrogram and heatmap of top-performed subjects #19 and #28. The left figure is the spectrogram, and the right figure is the attention heatmap. (a) Subject 19. (b) Subject 28.

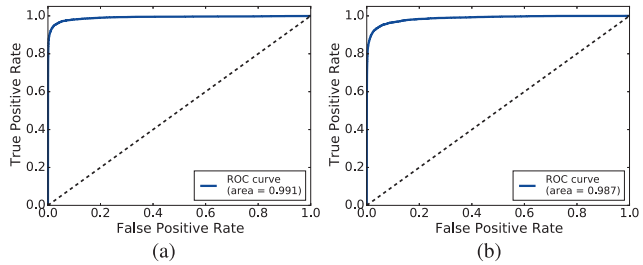


Fig. 13. ROC curves of the proposed model before and after applying the spectrotemporal mask. The black dashed line represents a nondiscriminatory test, where AUC equals 0.5. (a) Before applying mask. (b) After applying mask.

## V. Comparative Study

The goal of this comparative study is to highlight the superiority of our proposed CNN model over the commonly used classic and DL models when training on Optimal Gestures and Optimal Sensor derived in Section III-B. Thus, we compare our proposed model with: 1) a two-layer multilayer perceptron (MLP) model; 2) a two-module hybrid model with four CNN blocks followed by six long short-term memory (LSTM) layers and an FC layer; and 3) a classic SVM model. In this comparative study, each comparing model trains on Optimal Gestures (4, 12, 15, 22, 26, 30, and 32) and the most informative Optimal Sensors (6, 7, 8, 11, and 12). The validation data include sEMG signals from the even repetitions (2, 4, and 6) of the remaining 33 gestures, while the test data contain the sEMG signals from the odd repetitions (1, 3, and 5) of the same 33 gestures.

In this section, we select the comparing models (neural networks) to be comparable in terms of complexity to our proposed CNN model. The MLP model has 30 neurons on the hidden layer. We modify our recently proposed hybrid model [56] specifically for the identification problem. The hybrid model has a CNN module followed by an LSTM module. The CNN module consists of four CNN blocks, each having a 2-D convolutional layer, a batch normalization layer, and a ReLU layer. The convolutional layers have 20, 40, 60, and 80 channels, respectively, with a kernel size of $3 \times 3$. The output from the last CNN block is fed to the LSTM module,

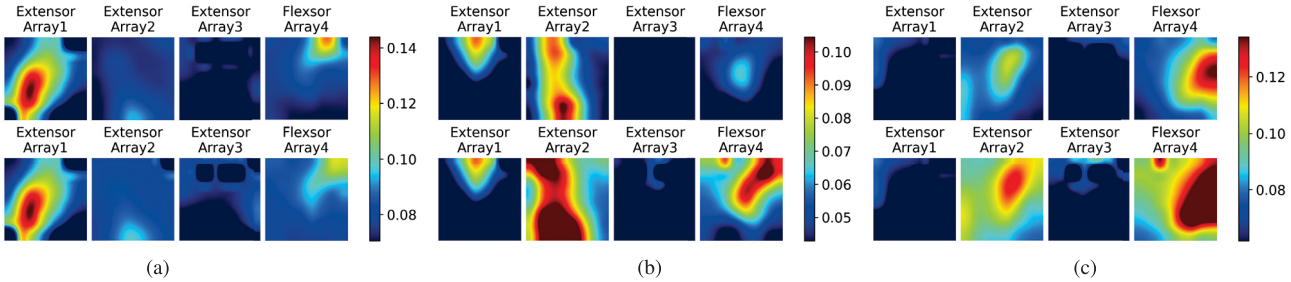


Fig. 14. Root mean square of the four 8 × 8 electrode arrays examples of the top-3-performing gestures. Arrays 1 and 3 were placed close to the wrist. Arrays 2 and 4 were positioned close to the elbow. For all subplots, the first row shows the HD-sEMG signals collected on Day 1, while the second row includes HD-sEMG signals on Day 2. (a) Middle finger extension. (b) Close hand. (c) Open hand.

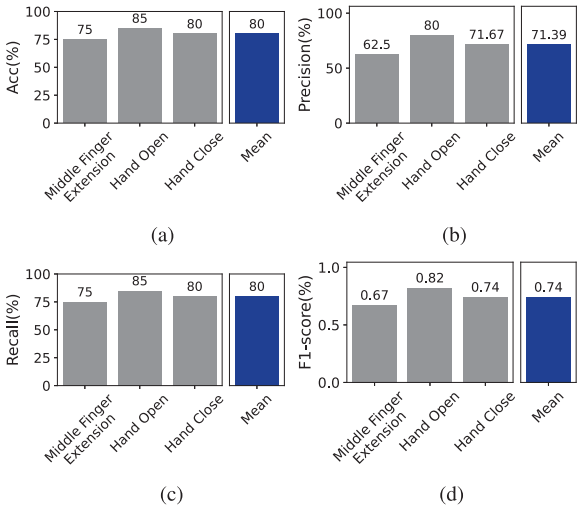


Fig. 15. Individual and averaged test accuracy, precision, recall, and F1 score of the top-3-performing gestures. (a) Average test accuracy. (b) Average precision. (c) Average recall. (d) Average F1 score.

which consists of six LSTM layers with 92 hidden units on each layer. The last layer is an FC layer with 4240 neurons. For the SVM model, we extract mean, median, root mean square, and variance from sliding windows of size 10 × 25 with a 20% overlap along the frequency axis on each input spectrogram. This results in a feature vector of size 124 for each sensor. This procedure is done for all five Optimal Sensors. Therefore, each sample spectrogram is converted to a vector of 620 features, which are reduced to seven principal components using PCA before feeding to the SVM. The results are summarized in Table III and highlighted in the following contributions.

*Observation 1:* The SVM fails in the identifying task. The hybrid model achieves about 79% accuracy, which is 8% lower than our proposed CNN.

*Observation 2:* The MLP achieves about 13% less average accuracy than the proposed model. The comparing MLP model architecture has to be simple (two layers) to match the structural complexity of our proposed model. Rather than flattening the inputs, training a CNN model preserves the spatiotemporal information of the spectrograms. Leveraging kernel sliding, our CNN model can detect neural feature patterns appearing anywhere in a spectrogram based on a smaller amount of training data than the data needed for training an MLP for the same task.

*Observation 3:* Our CNN model trains in 26% (5 s/19 s) of the time required by the hybrid model on each iteration. Given the similar model convergence, which is the number of training iterations for a model to achieve its maximum performance, our proposed model is more efficient.

Considering the above observations, the proposed model proves to be considerably well-suited with a compact, practical, explainable, and robust design for a PIS.

*Remark 2:* Gesture optimization in Section III-B1 is model-independent because it is the result of GMM clustering based on the extracted features of each gesture. However, sensor optimization based on performance ranking in Section III-B2 is sensitive to model types. To enhance the fairness of the comparison between our model and the compared ML/DL models, given the Optimal Gestures, we compare our proposed model with the abovementioned compared models on all 12 sensors. The comparison results show that our model can achieve 90.5% average accuracy when given full information from all the sensors, outperforming the hybrid, MLP, and SVM models by 0.9%, 10.2%, and 43.5%, respectively. As a result, in the next step, we conduct sensor optimization on the compared hybrid model because its identification performance on all 12 sensors is close to that of our proposed model. The sensor optimization on the two-module hybrid model results for the five Optimal Sensors (i.e., sensors 1, 6, 10, 11, and 12), which are not only less practical by having sensor placement at three locations on the forearm but also give a lower accuracy of 85.53% compared to our model's 87.75%.

*Remark 3:* To evaluate the performance of our proposed model over the conventional method for multisession biometric identification based on sEMG, we compare our model with SVM (used in [30]). We follow the same experiments in terms of preprocessing, windowing, and feature extraction as when comparing our model to SVM for single-session evaluation (the preprocessing steps can be found in Section III-E and the windowing and feature extraction mentioned in this section). To form a fair comparison, we identify the same 20 subjects through the most reliable gestures (i.e., middle finger extension, hand close, and hand open) described in Section IV-D. The total number of features extracted from each spectrogram input is 1984 (31 windows * four feature types * 16 sensors), which is further reduced to 20 (which explains 90% variances of the original feature space) after PCA to reduce computational efforts and avoid overfitting. As a result, the compared SVM model can identify 20 subjects with average accuracies of 58.89%, 57.01%, and 44.61% in middle finger extension, hand close, and hand open, respectively.

TABLE III
RESULTS FOR COMPARING THE PROPOSED MODEL WITH COMMONLY USED CLASSIC AND DL MODELS

| Models | # Trainable Parameters | Accuracy | Time/Epoch |
|---|---|---|---|
| **Proposed CNN** | **937,540** | **87.753%** | **5s** |
| Two-module Hybrid | 932,456 | 79.09% | 19s |
| Two-layer MLP | 938,770 | 74.97% | 2s |
| SVM | N/A | 44.116% | N/A |

Note: #: Number. s: second. N/A: Not Applicable.

Considering the performance of the proposed model in these scenarios (i.e., 75%, 85%, and 80%), this means that our proposed model outperforms the conventional method, SVM, by 16.11%, 27.99%, and 35.39%.

## VI. CONCLUSION

In this article, we investigate the possibility of using the hidden underlying neurophysiological patterns in multichannel sEMG signals to identify users while securing high performance. We propose and evaluate an optimized and explainable neural network that analyzes the information context of gestures and sensors to find out the minimum but sufficient number of Optimal Gestures, Optimal Sensors, and best frequency bands for training the model to enhance practicality and efficiency. We have also shown that the performance can be preserved using data from only two muscle groups. This article, for the first time, aims to tackle gesture-independent personal identification, demonstrating the capability of our model in extracting common, user-specific neurophysiological patterns across gestures. The Grad-CAM analysis is also performed to decode the attention of the neural network model. The outcome of Grad-CAM analysis is also utilized to reduce the needed data size and thereby reduce the number of trainable parameters of the model, reducing the complexity and increasing the speed of training. As a result of gesture and sensor optimization and Grad-CAM analysis, the proposed method can identify 40 subjects based on only 4% of training data from the database. The comprehensive evaluation of the proposed model on: 1) a multisession dataset using HD-sEMG and 2) a single-session dataset using bipolar sEMG, not only shows the robustness of the proposed method in generalization over time but also highlights the performance of the system under various experimental conditions, and experimental setups and sEMG modalities. It is worth noting that none of our methods (GMM clustering, CNN model structure, and Grad-CAM XAI) is completely new. They have been researched separately in other domains but not collectively in the domain of sEMG signal processing for personal identification purposes over the last two decades. This article sheds light on the capacity of the underlying neurophysiological signature of sEMG biosignals for identifying individuals. XAI helps visualize the unique and complex neural feature patterns associated with each subject and quantify these patterns through identification codes, pushing forward biometric research on human identification.

This article preliminarily proves that the proposed model is generalizable and robust in identifying 60 subjects (40 for single day and 20 for multiday biometric identification) from two datasets under different experimental settings without modifying the architecture. To further enhance the real-life practicality of sEMG-based biometric identification when considering the accelerated interest in using biosignals for identification, the future work in this research field can be: 1) raising the limit on the number of subjects by pooling multiple datasets to generate a multicenter benchmarking database to enhance system generalization; 2) collecting multiday signals from more subjects performing more gestures to enhance the flexibility and robustness of multiday identification; and 3) evaluating the system performance on recognizing intruders (i.e., unknown subjects to the system) through leave-one-subject-out cross validation to enhance system reliability and unbiasedness.

## REFERENCES


[1] C.-Y. Chou, E.-J. Chang, H.-T. Li, and A.-Y. Wu, "Low-complexity privacy-preserving compressive analysis using subspace-based dictionary for ECG telemonitoring system," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 4, pp. 801–811, Aug. 2018.

[2] T. Yaqoob, H. Abbas, and N. Shafqat, "Integrated security, safety, and privacy risk assessment framework for medical devices," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1752–1761, Jun. 2020.

[3] A. Abbas and S. U. Khan, "A review on the state-of-the-art privacy-preserving approaches in the e-health clouds," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 4, pp. 1431–1441, Jul. 2014.

[4] A. Khedr and G. Gulak, "SecureMed: Secure medical computation using GPU-accelerated homomorphic encryption scheme," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 2, pp. 597–606, Mar. 2018.

[5] R. Sánchez-Guerrero, F. A. Mendoza, D. Díaz-Sánchez, P. A. Cabarcos, and A. M. López, "Collaborative eHealth meets security: Privacy-enhancing patient profile management," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1741–1749, Nov. 2017.

[6] S. Cheng, J. Wang, D. Sheng, and Y. Chen, "Identification with your mind: A hybrid BCI-based authentication approach for anti-shoulder-surfing attacks using EEG and eye movement data," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[7] W. Yang, N. Li, O. Chowdhury, A. Xiong, and R. W. Proctor, "An empirical study of mnemonic sentence-based password generation strategies," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Vienna, Austria, Oct. 2016, pp. 1216–1229.

[8] M. Cardaioli, M. Conti, K. Balagani, and P. Gasti, "Your pin sounds good! augmentation of pin guessing strategies via audio leakage," in *Proc. Eur. Symp. Res. Comput. Secur.* Guildford, U.K.: Springer, Sep. 2020, pp. 720–735.

[9] S. Mamonov and R. Benbunan-Fich, "The impact of information security threat awareness on privacy-protective behaviors," *Comput. Hum. Behav.*, vol. 83, pp. 32–44, Jun. 2018.

[10] K. Erickson and P. N. Howard, "A case of mistaken identity? News accounts of hacker, consumer, and organizational responsibility for compromised digital records," *J. Comput.-Mediated Commun.*, vol. 12, no. 4, pp. 1229–1247, Jul. 2007.

[11] S. Alrwais et al., "Catching predators at watering holes: Finding and understanding strategically compromised websites," in *Proc. 32nd Annu. Conf. Comput. Secur. Appl.*, Dec. 2016, pp. 153–166.

[12] S. Hadiyoso, S. Aulia, and A. Rizal, "One-lead electrocardiogram for biometric authentication using time series analysis and support vector machine," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 1–14, 2019.

[13] Y. Zhang and M. Juhola, "On biometrics with eye movements," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1360–1366, Sep. 2017.

[14] P. Hu, H. Ning, T. Qiu, H. Song, Y. Wang, and X. Yao, "Security and privacy preservation scheme of face identification and resolution framework using fog computing in Internet of Things," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1143–1155, Oct. 2017.

[15] M. P. Yankov, M. A. Olsen, M. B. Stegmann, S. S. Christensen, and S. Forchhammer, "Fingerprint entropy and identification capacity estimation based on pixel-level generative modelling," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 56–65, 2019.

[16] L. Lu, J. Mao, W. Wang, G. Ding, and Z. Zhang, "A study of personal recognition method based on EMG signal," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 4, pp. 681–691, Aug. 2020.

[17] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial 'gummy' fingers on fingerprint systems," *Proc. SPIE*, vol. 4677, pp. 275–289, Apr. 2002.

[18] V. Ruiz-Albacete, P. Tome-Gonzalez, F. Alonso-Fernandez, J. Galbally, J. Fierrez, and J. Ortega-Garcia, "Direct attacks using fake images in iris verification," in *Proc. Eur. Workshop Biometrics Identity Manage.* Cham, Switzerland: Springer, 2008, pp. 181–190.

[19] S. K. Cherupally et al., "ECG authentication hardware design with low-power signal processing and neural network optimization with low precision and structured compression," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 198–208, Apr. 2020.

[20] C. De Luca, "Electromyography," in *Encyclopedia of Medical Devices and Instrumentation*. Hoboken, NJ, USA: Wiley, 2006.

[21] S. A. Raurale, J. McAllister, and J. M. D. Rincon, "Real-time embedded EMG signal analysis for wrist-hand pose identification," *IEEE Trans. Signal Process.*, vol. 68, pp. 2713–2723, 2020.

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

[23] T. Kurogi, H. Yamaba, K. Aburada, T. Katayama, M. Park, and N. Okazaki, "A study on a user identification method using dynamic time warping to realize an authentication system by s-EMG," in *Advances in Internet, Data & Web Technologies*. Cham, Switzerland: Springer, 2018, pp. 889–900.

[24] J. Fan et al., "Cancelable HD-SEMG biometric identification via deep feature learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1782–1793, Apr. 2022.

[25] W. Li, Z. Zhang, B. Hou, and A. Song, "Collaborative-set measurement for ECG-based human identification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–8, 2021.

[26] C. Tan, L. Zhang, T. Qian, S. Brás, and A. J. Pinho, "Statistical n-best AFD-based sparse representation for ECG biometric identification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[27] X. Jiang et al., "Cancelable HD-sEMG-Based biometrics for cross-application discrepant personal identification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 4, pp. 1070–1079, Apr. 2021.

[28] S. Shin, J. Jung, and Y. T. Kim, "A study of an EMG-based authentication algorithm using an artificial neural network," in *Proc. IEEE Sensors*, Oct. 2017, pp. 1–3.

[29] Q. Li, P. Dong, and J. Zheng, "Enhancing the security of pattern unlock with surface EMG-based biometrics," *Appl. Sci.*, vol. 10, no. 2, p. 541, Jan. 2020.

[30] X. Jiang et al., "Measuring neuromuscular electrophysiological activities to decode HD-sEMG biometrics for cross-application discrepant personal identification with unknown identities," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[31] R. Shioji, S.-i. Ito, M. Ito, and M. Fukumi, "Personal authentication and hand motion recognition based on wrist EMG analysis by a convolutional neural network," in *Proc. IEEE Int. Conf. IoT Intell. Syst. (IOTAIS)*, Nov. 2018, pp. 184–188.

[32] J. He and N. Jiang, "Biometric from surface electromyogram (sEMG): Feasibility of user verification and identification based on gesture recognition," *Front. Bioeng. Biotechnol.*, vol. 8, p. 58, Feb. 2020.

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

[34] *BioRender*. Accessed: Jan. 28, 2022. [Online]. Available: https://biorender.com/

[35] G. L. Iverson, "Z scores," in *Encyclopedia of Clinical Neuropsychology*, J. S. Kreutzer, J. DeLuca, and B. Caplan, Eds. New York, NY, USA: Springer, 2011, pp. 2739–2740.

[36] E. Rahimian, S. Zabihi, S. F. Atashzar, A. Asif, and A. Mohammadi, "Xceptiontime: Independent time-window xceptiontime architecture for hand gesture classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 1304–1308.

[37] E. Rahimian, S. Zabihi, A. Asif, S. F. Atashzar, and A. Mohammadi, "Few-shot learning for decoding surface electromyography for hand gesture recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1300–1304.

[38] E. Tyacke et al., "Hand gesture recognition via transient sEMG using transfer learning of dilated efficient CapsNet: Towards generalization for neurorobotics," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9216–9223, Oct. 2022.

[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 818–833.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-Resnet and the impact of residual connections on learning," in *Proc. 31th AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 1–7.

[41] G. Jia, H.-K. Lam, S. Ma, Z. Yang, Y. Xu, and B. Xiao, "Classification of electromyographic hand gesture signals using modified fuzzy C-means clustering and two-step machine learning approach," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1428–1435, Jun. 2020.

[42] M. K. Burns, D. Pei, and R. Vinjamuri, "Myoelectric control of a soft hand exoskeleton using kinematic synergies," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1351–1361, Dec. 2019.

[43] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, *Introduction to EEG-and Speech-Based Emotion Recognition*. New York, NY, USA: Academic, 2016.

[44] K. P. Thomas and A. P. Vinod, "Toward EEG-based biometric systems: The great potential of brain-wave-based biometrics," *IEEE Syst. Man, Cybern. Mag.*, vol. 3, no. 4, pp. 6–15, Oct. 2017.

[45] S. Wang, G. Azzari, and D. B. Lobell, "Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques," *Remote Sens. Environ.*, vol. 222, pp. 303–317, Mar. 2019.

[46] J. Wang and J. Jiang, "Unsupervised deep clustering via adaptive GMM modeling and optimization," *Neurocomputing*, vol. 433, pp. 199–211, Apr. 2021.

[47] X. Zhou, G. Bian, X. Xie, Z. Hou, X. Qu, and S. Guan, "Analysis of interventionalists' natural behaviors for recognizing motion patterns of endovascular tools during percutaneous coronary interventions," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 2, pp. 330–342, Apr. 2019.

[48] J. Chen, G. Wang, and G. B. Giannakis, "Nonlinear dimensionality reduction for discriminative analytics of multiple datasets," *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 740–752, Feb. 2019.

[49] M. Nishida and T. Kawahara, "Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, Apr. 2003, p. 1.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[51] P. Virtanen, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[52] S. Benatti, E. Farella, E. Gruppioni, and L. Benini, "Analysis of robust implementation of an EMG pattern recognition based control," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process.*, Mar. 2014, pp. 1–10.

[53] X. Jiang et al., "Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1035–1046, 2021.

[54] D. Biswas et al., "CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 2, pp. 282–291, Apr. 2019.

[55] B. Taşar, "Deep-BBiIdNet: Behavioral biometric identification method using forearm electromyography signal," *Arabian J. Sci. Eng.*, vol. 47, no. 11, pp. 14571–14581, Nov. 2022.

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