Reduce System Redundancy and Optimize Sensor Disposition for EMG–IMU Multimodal Fusion Human–Machine Interfaces With XAI

Peiqi Kang, *Student Member, IEEE*, Jinxuan Li, *Student Member, IEEE*, Shuo Jiang, *Member, IEEE*, and Peter B. Shull, *Member, IEEE*

***Abstract*— Multimodal sensor fusion can improve the performance of human–machine interfaces (HMIs). However, increased sensing modalities and sensor count often cause excess redundancies, and when applying deep learning approaches, the recognition system can become overly complex and difficult for humans to understand. In this article, we propose an explainable artificial intelligence (XAI) approach to reduce redundancies in inertial measurement units (IMUs) and electromyography (EMG) multimodal systems and optimize sensor disposition in prosthetic hand control. Four attribution algorithms and four quantitative evaluation algorithms were used on an open-source dataset of 17 hand gestures from 60 healthy subjects and 11 amputees to explore the working mechanism behind the multimodal system. Using an XAI approach, we reduced the total number of required sensors by 40% while maintaining the same level of accuracy. These results could enable optimized HMI system design with reduced sensor costs and manufacturing costs. The proposed approach lays the foundation for improving HMI systems by reducing complexity and revealing explainable information that is typically hidden within deep neural networks, thereby facilitating patients in the daily use of prosthetic hands and helping improve their quality of life.**

***Index Terms*— Explainable artificial intelligence (XAI), hand gesture recognition, human–machine interfaces (HMIs), multimodal sensor fusion.**

I. Introduction

ELECTROMYOGRAPHY (EMG) is an important human–machine interface (HMI) technology that has significant medical applications, including prosthetic control and stroke rehabilitation [1], [2]. However, EMG still suffers from inherent limitations, including nonstationarity, low robustness, and low resolution to complex motions [3]. Therefore, in recent years, sensor fusion technology has been widely deployed in EMG-based HMI and greatly improved hand gesture recognition accuracy [4], [5].

However, the tradeoff for performance improvement results in a larger system with a more complex structure and higher cost [6]. Therefore, understanding the mechanism of how each sensing modality contributes to the system performance is vital to reducing the HMI system's redundancy [7]. With the popularity of deep learning, recognition models are becoming more complex, less explainable, and harder for humans to understand [8]. The multimodal HMI and recognition models are also becoming unexplainable black boxes [9]. The unexplainable models not only bring difficulties to the optimization of system design but also make it hard for society to trust these systems, and it could cause potential problems in responsibility identification and data-based decision-making [10], [11].

Recently, some representative studies on explainable artificial intelligence (XAI) have been proposed [12]. These studies clarified basic concepts and definitions of XAI and provided precursory demonstrations in the fields of medical image processing [13], autonomous driving [14], and natural language processing [15]. However, the explainability of multimodal HMI is seldom investigated by academics and industries. Since the contribution of each modality and each sensor unit is unclear in an unexplainable black box, the working mechanisms are unclear, leading to an increase in the system's complexity and manufacturing cost, and also increasing power consumption and decreasing endurance. In addition, in current hand gesture recognition systems, improvement is still based on experiments and experience; with an explainable system, the working mechanisms will be more intuitive, which will help researchers to provide more accurate and helpful solutions [16].

In this article, to bridge the gap between interpretability and high performance, we focused on two vital topics in the HMI system: reducing multimodal system redundancy and improving the performance of prosthetic hand control with a less complex sensing system. First, we quantitatively investigated the contribution of different sensing modalities on different muscle positions and verified the XAI explanations' rationality with physiology explanations. Second, we analyzed the performance of the EMG-ACC multimodal system on amputees to reveal the influence of sensing modalities and sensing positions on recognition performance and proposed an optimized sensor disposition with significantly reduced redundancy. Third, we utilized four quantitative evaluation methods to comprehensively assess the explanation results generated by XAI on faithfulness, sensitivity, complexity, and randomization metrics

Manuscript received 18 August 2022; revised 6 November 2022; accepted 10 December 2022. Date of publication 26 December 2022; date of current version 10 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 52250610217 and Grant 52105033, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by the Chenguang Program by Shanghai Municipal Education Commission under Grant 21CGA23. The Associate Editor coordinating the review process was Dr. Shovan Barma. *(Corresponding authors: Shuo Jiang; Peter B. Shull.)*

Peiqi Kang, Jinxuan Li, and Peter B. Shull are with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: pekkykang@sjtu.edu.cn; jinxli@sjtu.edu.cn; pshull@sjtu.edu.cn).

Shuo Jiang is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Frontiers Science Center for Intelligent Autonomous Systems, Shanghai 200120, China (e-mail: jiangshuo@tongji.edu.cn).

Digital Object Identifier 10.1109/TIM.2022.3232159

to help researchers to select a more accurate explanation result.

To our best knowledge, this is the first study that focuses on explainable hand gesture recognition or prosthetic hand control sensor fusion algorithms. The proposed method successfully reduced the system complexity and improved the performance. Our work could provide urgently needed information to guide the design of multimodal HMI. It could also help patients to understand the treatment they receive and reduce their insecurity while helping doctors explain their decision to implement AI and reduce potential ethical and legal risks.

II. Background Research

A. Sensor-Fusion-Based HMI

In real-life scenarios, since the sensing targets are usually too complex for a single sensing modality to capture enough information to support the recognition model's decision, multisensor fusion solutions were proposed [17], [18]. For instance, in autonomous vehicles, to fully grasp complex road information, researchers always adopt sensor fusion solutions, including lidar, millimeter-wave radar, and vision cameras [19]. In biometrics areas, since muscle activities contain multidimensional information, with integrated multimodal systems (e.g., near-infrared sensors for collecting muscle hemodynamics information [20], the ultrasonic imaging for collecting muscle morphological information [21], and EMG for collecting muscle electricity [22]), recognition accuracy for a single task can be improved [6]. Jiang et al. [23] proposed a force myography (FMG)–EMG sensor fusion wristband for hand gesture recognition, which overcomes the sweat vulnerability of EMG and the low information density of FMG. Ceolini et al. [24] proposed a camera-EMG fusion hand gesture recognition system, which aims to utilize visual information to improve recognition accuracy or recognize objects during grasping to adjust the prosthetic. Krasoulis et al. [25] proposed the inertial measurement unit and electromyography (IMU–EMG) sensor fusion hand gesture recognition method, which successfully improved the recognition accuracy of amputees from 40% to around 80%.

For sensor fusion algorithms, there are four categories: pixel (raw data) level fusion, feature level fusion, decision level fusion, and hybrid fusion [26], [27]. Among these, the pixel level and the feature level are the two most common categories, and the most representative method is a matrix-based fusion method called multimodal tensor fusion network (TFN) [28]. However, since the TFN method requires lots of tensor outer product during the fusion process, which requires huge computation resources and lacks high-order fusion ability, modified versions of this method such as row-rank multimodal fusion [29] and polynomial tensor pooling [30] were also proposed. Other sensor fusion algorithms based on attention mechanism [31], adversarial learning [32], and autoencoder [33] were also applied to natural language processing and image processing applications. In addition, with the rapid development of sensor fusion technologies, how to select and evaluate these methods has become a key problem.

B. Explainable Artificial Intelligence

With the rapid development of deep learning technology, some models' parameters have exceeded the order of magnitude of millions. These complex models are impossible for humans to understand. Therefore, developing explainable methods becomes extremely important. There are mainly two kinds of explanation methods: gradient-based and perturbation-based. The gradient-based methods analyze the gradient flow through a model to generate explanations. For example, the layer conductance based on integrated gradients and their flow through the hidden neuron can provide researchers with pictures of neuron importance of a neural network [34]. The DeepLIFT method based on back-propagation and the gradient SHAP method based on Shapley values proposed in cooperative game theory can provide researchers with images of primary attribution of input data [35], [36]. The perturbation-based methods perturb input values and measure the change in the model's output to generate explanations. Occlusion and Shapley value sampling are both perturbation-based methods that can generate attribution for input data [37], [38]. Occlusion replaces each input feature with a given value to analyze the difference in output, and Shapley value sampling adds the feature values to a baseline to analyze the difference in output.

In addition, since XAI has drawn great attention from academia and various XAI methods have been proposed in recent years, choosing appropriate XAI methods by reasonable quantitative evaluation methods also becomes vital for success and correct explanation results. In general, XAI methods can be evaluated on five aspects: faithfulness, sensitivity, complexity, randomization, and localization [39]. Faithfulness evaluates whether important features also play important roles in the prediction process. Sensitivity quantifies whether the explanations are stable when encountering slight perturbations. Complexity evaluates the concise degree of explanations. Randomization can quantify how explanations deteriorate when the network's parameters become randomized. Finally, localization tests if the explanation methods are concentrated on the target object. In different scenarios, these five quantitative evaluation methods are not equally important, and not all methods are necessary at the same time.

Machine learning and deep learning have improved the human ability to diagnose and treat diseases [40]. However, due to the ethical and legal requirements, doctors must explain the output results of these models, and patients must understand these results [41]. Therefore, XAI is extremely important to medical applications. Molle et al. [42] proposed visualized convolutional neural networks (CNNs) to support skin lesion classification. Biffi et al. [43] proposed interpretable anatomical feature learning methods through deep generative models to provide an explanation for deep-learning-based cardiac remodeling. Jin et al. [44] provided a comprehensive evaluation of multimodal medical imaging and proposed a modality-specific feature importance metric to encode the clinical requirements on modality prioritization and feature localization. To our best knowledge, there are few related studies on explainable hand gesture recognition or prosthetic hand control sensor fusion algorithms. Our work could provide urgently needed information to the community and could guide the design of future wearable sensor fusion systems and algorithms.

III. Methods

A. Target Model

This article tries to use XAI methods to obtain the performance variation mechanism of different sensing modalities

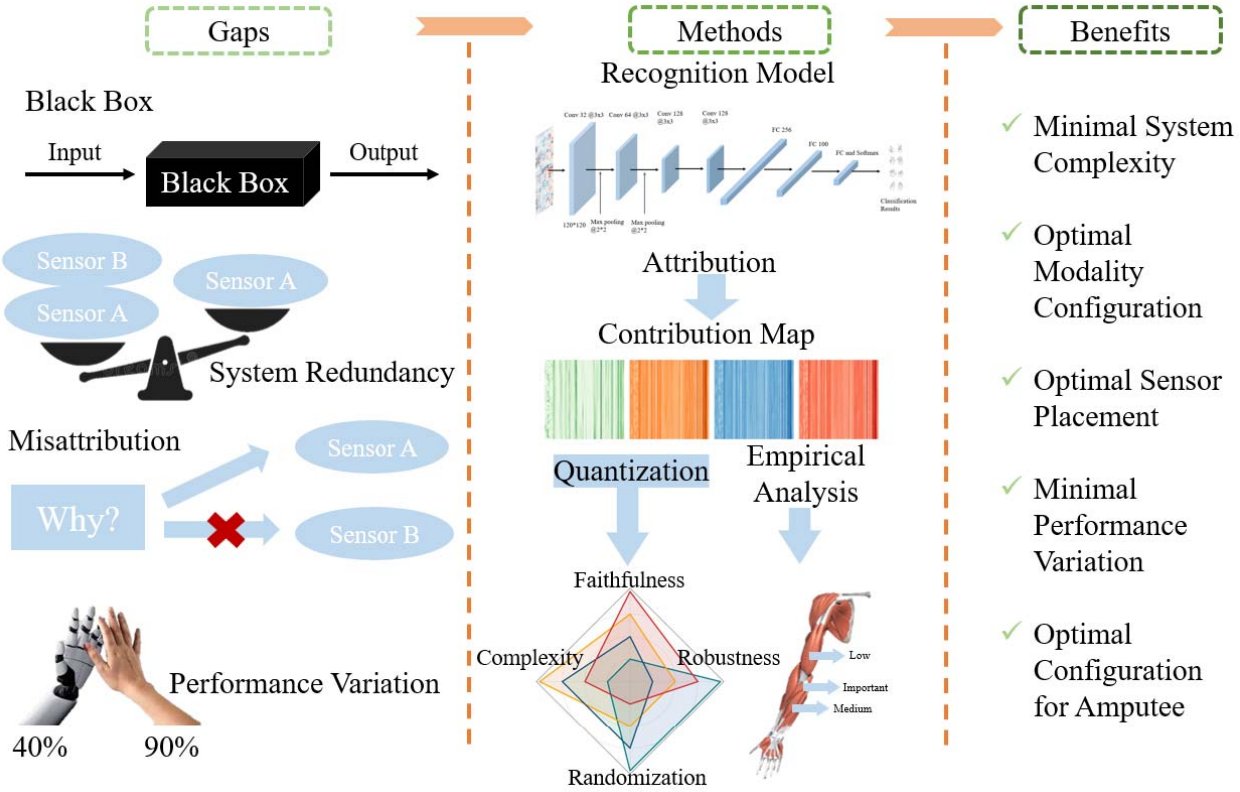


Fig. 1. Architecture of the proposed methods. This article proposed an analysis methodology that includes state-of-the-art XAI algorithms and empirical knowledge to reduce the system's complexity and provide an optimized sensor disposition.

and dispositions (see Fig. 1) and, based on the results, proposes a redundancy reduction and disposition solution. First, a target model needs to be built for analysis. In general, for hand gesture recognition, there are two main kinds of recognition algorithms: traditional machine learning models and deep learning models. The deep learning models have surpassed the machine learning models in many aspects, and in this article, we built a representative time-series recognition model as the target model (Fig. 2). The architecture of the model, from top to bottom, is a 1-D convolutional layer with 32 channels (kernel size is three), a 1-D convolutional layer with 64 channels (kernel size is three), two 1-D convolutional layers with 128 channels (kernel size is one), and three full collected layers (units are 256, 100, and 17). The input is 60 ms of EMG, accelerometer (ACC), gyroscope (GYRO), and magnetometer (MAG) signal of 120 channels, and the output is the hand gesture recognition result. To validate the reasonability of using the neural network as the target model, we tested it with datasets introduced in Section IV. When tested with an EMG–IMU multimodal fusion dataset, the target model's recognition accuracy of the healthy group is 95.6%. When tested with the EMG–ACC fusion dataset, the model's recognition accuracy of the healthy group is 91.5% and amputees is 79.2%.

B. Explanation Generation

For the CNN, the contribution of each data point to the final recognition can be represented or indirectly expressed as the gradient (or other parameters) of the model. Therefore, after the target model was built, attribution algorithms were utilized to generate the contribution map. These attribution algorithms can calculate the contribution distribution of input data to the

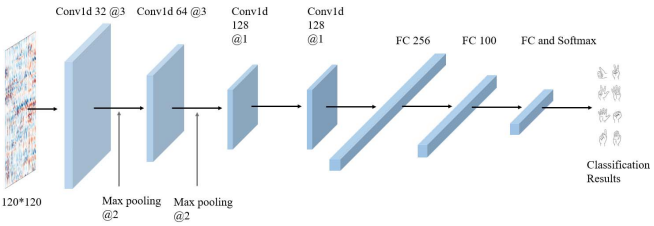


Fig. 2. Network architecture of the target model.

recognition model. By analyzing the contribution distribution combined with sensor modality and placement information, the performance variation mechanism of different sensing modalities and dispositions can be acquired. Since different attribution algorithms have different results, to provide a more comprehensive explanation of results, four state-of-the-art attribution algorithms were deployed.

Saliency is a classic method to calculate the input data's contribution attribution [45]. The basic idea of the saliency method is to process the first-order Taylor expansion for the model input and take its weight coefficient as the gradient and its absolute value as the contribution to the classification result

$$S_f(x_i) \approx w^T x_i + b \tag{1}$$

$$w = \frac{\partial S_f}{\partial x_i}_{i_0} \tag{2}$$

where S_f is the saliency score function of a neural network and w is the derivative of S_f at point i_0. Because the saliency method was proposed years ago, it may not outperform other XAI methods in any dimension. However, the saliency method is a classic method and has become the foundation of many

other XAI methods. In addition, the working principle of the saliency method is intuitive and easy to understand. Therefore, we think including the saliency could provide readers with more complete information.

Input X gradient is an extension of the saliency approach [46]. Different from the saliency method directly using the gradients as the contribution, input X gradient takes the gradients and multiplies by the input feature values as the total contribution of the input to the classification result.

Integrated gradient (IG) is the most representative XAI method [47]. For IG, the attribution of the ith input is defined as the integral of gradients from a given baseline to input along a straight path

$$\mathrm{IG}_i(x) ::= \left(x_i - x_i{}'\right) \times \int_{\alpha=0}^{1} \frac{\partial f\left(x' + \alpha \times \left(x - x'\right)\right)}{\partial x_i} d\alpha \quad (3)$$

where $x_i{}'$ is the baseline, x_i is the i input, f is the target network for explanation, and $(\partial f(x)/\partial x_i)$ is the gradient of f on the ith dimension.

The gradient SHAP (Shapley additive explanation) combines integrated gradients and SHAP into a single expected value equation [36]. For each prediction, the model generates a prediction value that is the summation of the SHAP value of each input

$$y_i = y_{\text{base}} + \mathrm{Sh}(x_{i1}) + \mathrm{Sh}(x_{i2}) + \cdots + \mathrm{Sh}(x_{\text{ik}}) \quad (4)$$

where x_i is the i input sample, the j feature of x_i is x_{ij}, the model's prediction value is y_i, and the baseline is y_{base}. The $\mathrm{Sh}(x_{\text{ij}})$ is the SHAP value of the x_{ij}. In general, the $\mathrm{Sh}(x_{\text{ij}})$ is the contribution of the x_{ij} to the prediction result.

C. Explanation Validation

As mentioned above, different attribution algorithms will provide different contribution results, and quantitatively evaluating the attribution results is vital for the accuracy of the explanation. This article gives full quantitative evaluations of the attribution algorithms from faithfulness, sensitivity, complexity, and randomization aspects. For the faithfulness assessment, we adopted the faithfulness correlation (FC) method [48]. By measuring the correlation between the randomly selected and baseline value replaced subset of given attributions and the difference in function output, the FC can provide the faithfulness source for a given XAI algorithm

$$\mathrm{Faith}(f, g; x) = \mathrm{corr}\left(\sum_{i \epsilon S} g(f, x)_i,\ f(x) - f(\bar{x}_s)\right) \quad (5)$$

where g is an explanation function, S is the randomly sampled subsets, $x_s = \{x_i, i \epsilon S\}$, and $\bar{x}_s$ is the input that is set to the baseline.

For the sensitivity assessment, we adopted the max-sensitivity (MS) method [49]. The MS method proposes to measure the maximum sensitivity of an explanation as follows:

$$\mathrm{SENS}_{\mathrm{MAX}}(g, f, x) = \max \| g(f(x + \epsilon)) - g(f(x)) \| \quad (6)$$

$$\left[\nabla_x g(f(x))\right]_j = \lim_{\epsilon \to 0} \frac{g\left(f\left(x + \epsilon e_j\right)\right) - g(f(x))}{\epsilon} \quad (7)$$

where e_j is the jth coordinate basis vector, whose jth entry is one and all others are zero.

For the complexity assessment, we adopted the sparseness method [50]. The sparseness value is calculated by the Gini Index, and a higher sparseness value (Gini Index) indicates a more concise explanation

$$G(v) = 1 - 2 \sum_{k=1}^{d} \frac{v_{(k)}}{\|v\|_1} \left(\frac{d - k + 0.5}{d}\right) \quad (8)$$

where v is a vector of nonnegative values and d is the dimension of v and $k \epsilon [d]$.

For the randomization assessment, we adopted the model parameter randomization (MPR) method [51]. The working principle of MPR is randomizing the parameters of single model layers and then measuring the difference between the new explanation and the original explanation.

All four XAI methods (saliency, integrated gradients, gradient SHAP, and input X gradient) were used to perform the attribution of input data and four quantitative evaluation methods were used to evaluate the attribution result. Based on the quantitative evaluation result, the best-performing XAI method's attribution result was chosen as the final result for further analysis. In addition, the consistency between the XAI results and empirical analysis is also important. Based on the knowledge of anatomy, a clear relationship between muscles and motions can be obtained. The working principles of different sensing modalities have also been clearly summarized [3]. The problem is that the empirical analysis is not quantized and can only be described by empirical statements. However, the results of XAI methods should not conflict with empirical analysis. Therefore, it is important for the two methods to be cross-validated, and in this article, we provide the XAI results with detailed physiological evidence.

IV. Experiments

A. Dataset

The open-source Nina pro database's second, third, and seventh subsets were chosen to validate our proposed methods [25], [52]. Database 2 (DB2) [52] included EMG and ACC data of 40 healthy participants. Database 3 (DB3) [52] included EMG and ACC data of 11 amputee participants. The purpose of DB2 and DB3 is to analyze the performance difference of sensor fusion technology between healthy participants and amputee participants. Database 7 (DB7) [25] included EMG, ACC, GYRO, and MAG data of 20 healthy participants. The purpose of DB7 is to provide a comprehensive analysis of the performance of different sensor fusion solutions. The EMG signals were sampled at 2 kHz, and the IMU (ACC, GYRO, and MAG) data were sampled at 128 Hz.

Seventeen gestures were included in these datasets. Eight of them are movements of the fingers, including: 1) thumb up (TU); 2) extension of index and middle, flexion of the others (EIM); 3) flexion of the ring and little finger, the extension of the others (FRL); 4) thumb opposing base of the little finger (TO); 5) abduction of all fingers (AA); 6) fingers flexed together in the first (FF); 7) pointing index (PI); and 8) adduction of extended fingers (AE). Nine of them are wrist movements: 9) wrist supination (axis: middle finger) (WSM); 10) wrist pronation (axis: middle finger) (WPM); 11) wrist supination (axis: little finger) (WSL); 12) wrist pronation (axis: little finger) (WPL); 13) wrist flexion (WF); 14) wrist extension (WE); 15) wrist radial deviation (WRD); 16) wrist ulnar deviation (WUD); and 17) wrist extension with closed hand (WEC).

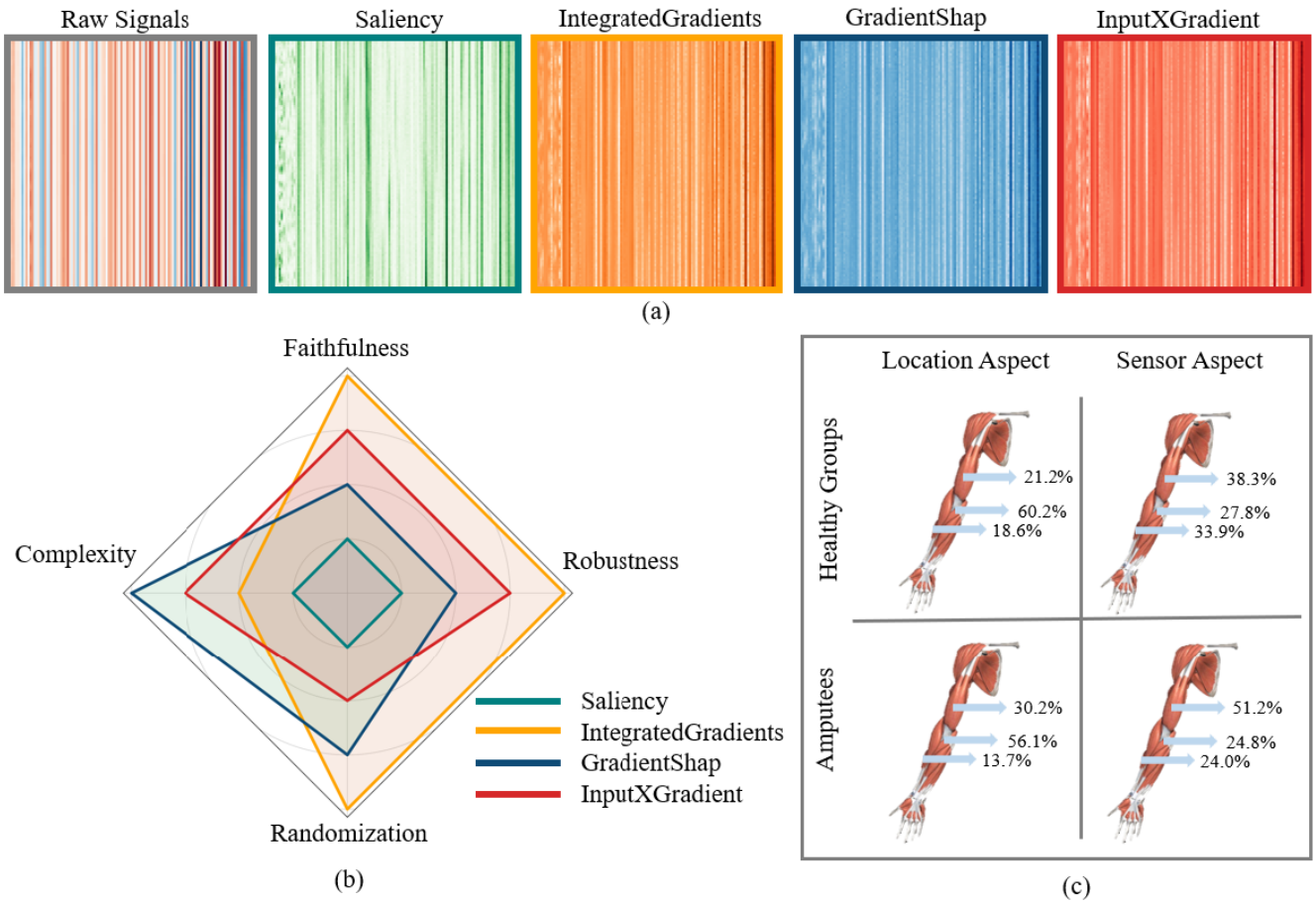


Fig. 3. (a) Attribution results of four XAI methods. (b) Quantitative evaluation results of four XAI methods. (c) Contribution difference of different sensor placement locations.

The datasets were collected by 12 fusion sensors placed on participants' arms. The first eight sensors were equally spaced around the forearm, and the rest were placed on the extensor digitorum communis muscle (EDC), flexor digitorum superficialis muscle (FDS), biceps brachii, and triceps brachii.

B. Experimental Protocol

This article separately conducted three experiments on two groups of participants (healthy groups and a group of amputees, six experiments in total). These experiments measured the contribution difference of four sensing modalities, the contribution difference of different gesture categories, and the influence of sensor placement location on sensor fusion performance.

For the healthy groups, 17 gestures of 20 healthy participants were collected by 12 fusion sensors (DB7). Each fusion sensor contained a single-channel EMG, a three-axis ACC, a three-axis GYRO, and a three-axis MAG. Since the hand gestures were static and according to previous work, the raw data were segmented into 60 ms per segment before being processed by the neural network [53], [54], [55]. Therefore, each input of the classification network was a 120×120 matrix (the sampling rate was 2 kHz, and 120 axes of four different modalities were used). After preprocessing, the target network was trained with the segmented data. According to the dataset's source paper, sixfold cross-validation was adopted in the training process [25]. When the target network was trained, we used four XAI algorithms introduced in Section III to generate the attribution map for further analysis.

For the amputee group, 17 gestures of 40 healthy participants and 11 amputee participants were collected by 12 fusion sensors (DB2 and DB3). Each fusion sensor contained a single-channel EMG and a three-axis ACC. Before being processed by the neural network, the raw data were segmented into 60 ms per segment. Therefore, each input of the classification network was a 120×48 matrix (the sampling rate was 2 kHz, and 48 axes of four different modalities were used). After preprocessing, the target network was trained separately with the segmented data of the healthy participants and amputees, and then it generated the attribution map of the 48-axis sensor.

V. Results and Discussion

A. Quantitative Evaluation of the XAI Methods

Since the attribution results of different XAI methods are not consistent, it is important to select the most suitable one for this problem. We adopted four quantitative methods introduced above to evaluate the four XAI methods on faithfulness, sensitivity, complexity, and randomization aspects. To ensure the stability of the evaluation, we took the average value across subjects and gestures. As shown in Fig. 3(b), the IG method showed the best performance in faithfulness, sensitivity, and randomization. Although it did not get the best performance in complexity, the complexity characteristics only represent the

TABLE I
CONTRIBUTION OF THE FOUR SENSING MODALITIES

Gestures	EMG	ACC	GYRO	MAG
TU	18.19%	27.88%	11.68%	42.25%
EIM	25.88%	23.19%	13.04%	37.89%
FRL	20.36%	32.01%	9.85%	37.78%
TO	18.86%	27.31%	9.94%	43.89%
AA	10.27%	30.97%	8.94%	49.82%
FF	21.26%	28.85%	10.69%	39.20%
PI	17.09%	28.96%	10.76%	43.19%
AE	18.66%	27.62%	6.53%	47.20%
WSM	12.84%	27.52%	8.49%	51.16%
WPM	17.36%	29.47%	11.36%	41.82%
WSL	16.09%	30.61%	17.44%	35.87%
WPL	20.70%	26.16%	10.54%	42.60%
WF	14.18%	25.34%	8.11%	52.37%
WE	15.57%	18.76%	8.02%	57.65%
WRD	20.01%	23.00%	10.90%	46.09%
WUD	17.65%	25.66%	10.02%	46.68%
WEC	14.24%	23.67%	10.28%	51.81%

concise degree of the XAI method. Therefore, we chose the IG method for further analysis.

B. Results of Explanations of Healthy Group

As shown in Fig. 3(a), the results of all four XAI methods showed prominent modality properties and axis properties, which affirms the rationality of choosing sensing modalities and sensor axes as analysis objects. By averaging the attribution of all axes of a sensing modality and then calculating the mean value among all the subjects and all the hand gestures, we found that the most important sensing modality is MAG, then ACC, EMG, and GYRO. MAG, ACC, EMG, and GYRO contributed 45.1%, 26.9%, 17.6%, and 10.4%, respectively, to the final recognition results. The detailed contribution of the four sensing modalities on 17 hand gestures is listed in Table I.

In addition, the hand gesture categories also showed impacts on the performance of sensing modalities. For the eight-finger movement gestures, MAG, ACC, EMG, and GYRO contributed 42.7%, 28.3%, 18.8%, and 10.2%, respectively, to the final recognition results. For the nine-wrist movement gestures, MAG, ACC, EMG, and GYRO contributed 47.3%, 25.6%, 16.5%, and 10.6%, respectively, to the final recognition results. In addition, from the attribution map (Fig. 4), EMG also showed a higher temporal resolution, indicating that it could capture fine-grained signals. These results showed that EMG is better at recognizing fine-grained gestures compared with recognizing directional gestures. It is worth mentioning that, although MAG showed a huge contribution to the recognition results, according to the attribution map (see Fig. 4), MAG sensor is strongly location-dependent. In a real-life scenario, the MAG might suffer a huge accuracy decrease when encountering placement location shifts.

GYRO took 30% of the system sensing channels but only contributed 10.4% to the recognition accuracy. Compared with

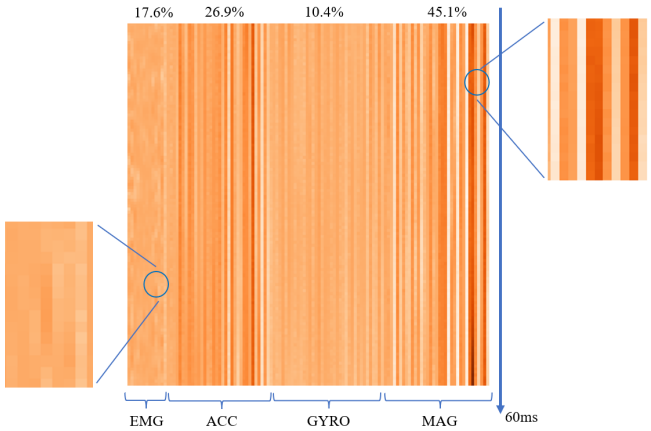


Fig. 4. Attribution map of the integrated gradients method. EMG showed higher temporal resolution; however, for the final recognition results, the classifier paid more attention to MAG. In addition, the placement location showed a great impact on the sensor performance, especially the MAG sensors.

ACC and MAG, which each took 30% of the sensing channels but contributed 45.1% and 26.9% to the recognition results, GYRO seriously increased the system's redundancy. Therefore, from the XAI results, to reduce the system's complexity, removing GYRO is the best option. Also, in actual use, GYROs are usually high power-consuming, so removing them will also improve the endurance of the system. If further simplification is required, then the target hand gestures need to be taken into consideration. If the gesture set contains many fine-grained gestures, then EMG should not be removed. Therefore, it is necessary to choose customized sensor solutions for specific tasks or to design specific gestures for different sensor hardware.

In addition, the conclusion from the XAI methods can be validated from the physiology aspect. Based on our previous systematic study on the working principle of different sensing modalities and their measured biological characteristics [3], EMG captures muscles' neural firing information (current intensity), which does not include any intuitional motion information or directional information. Therefore, EMG is sensitive to fine-grained gestures (e.g., finger movement gestures) but has difficulty recognizing gross gestures or directional gestures (e.g., wrist movement gestures). ACC can capture both the vibration of muscle contraction and the acceleration of motion signals. Therefore, it is suitable for both fine-grained gestures and gross gestures. MAG can directly measure the posture and recognize gestures from it. Therefore, the MAG has a high recognition accuracy for static gestures. While, in real-life applications, MAG frequently has serious drifting problems, hindering the MAG sensing method from achieving high performance. However, GYRO mainly captures the rotation signals which are mainly seen in the sign language gesture set. Therefore, for interaction gesture sets, the contribution of the GYRO sensor is relatively low.

Based on the sensor placement location provided by the dataset's introduction, 12 sensing locations were divided into three categories: Category A, including eight fusion sensors equally spaced around the forearm; Category B, including two fusion sensors placed on the EDC and FDS; and Category C, including two fusion sensors placed on the biceps brachii and triceps brachii. For the three placement locations

[Fig. 3(c) (left)], location A contributed most to the recognition results (60.2%), location C contributed 21.2%, and sensors at location B contributed least to the recognition results (18.6%). Also, the gesture categories did not show any impact on the locations' contribution distribution. For a single sensor (Fig. 3(c) right, all the results were averaged among sensor numbers and subjects), if it is placed on location A can contribute 27.8% to the final recognition results, 33.9% on location B, and 38.3% on location C. Similarly, the gesture categories did not show any impacts on the locations' contribution distribution. In addition, if taking sensor placement location into consideration, GYROs on location A only contribute 7.0% to the recognition results, on location B, 2.2%, and on location C, 2.6%. To be more specific, each GYRO sensor on locations A, B, and C only contributes 0.29%, 0.37%, and 0.43% to the recognition results. However, the average contribution of four sensor modalities in these three locations is 0.75%, 0.93%, and 1.06%. The contribution of GYRO is significantly lower than other sensing modalities.

Based on the quantitative results, we can see that the sensors placed on individual muscles have better performance, while sensors placed on the forearm muscles may experience interference from signal crosstalk. These results also explained that interaction-purposed devices are always made into band-shapes, which are easier to wear and more comfortable to use. The band shape will damage the performance but can meet daily requirements. However, for medical purposes, like prosthetic control, muscles including EDC, FDS, biceps brachii, and triceps brachii are indispensable.

C. Results of Explanations of Amputees

Prosthetic hand control is an important application of hand gesture recognition. Each of the amputees lost their hand and part of the forearm, leading to the amputees' motor ability being significantly lower than the healthy group's. In addition, due to much less motor stimulation, the amputees' muscles on the upper and lower arms are significantly atrophied. This means that many times, the amputees have the motion intention but lack actual movements. The lack of actual movements reduces the quality of the ACC signal and thus causes bad recognition results. For EMG, since muscle atrophy will lead to weak muscle activation intensity, the signal-to-noise ratio will be lower and thus will cause bad EMG signal quality and recognition accuracy. In addition, since it is hard to collect amputees' data, providing researchers with knowledge on collecting amputees' signals will help the development process of prosthetics. Compared with the healthy groups, the recognition of all the sensing modalities on amputees is significantly lower. The EMG-ACC fusion system's recognition accuracy dropped from 91.5% to 79.2%, and the ACC's recognition accuracy dropped from 90.0% to 76.8%.

By averaging the attribution of all axes of a sensing modality and then calculating the mean value among all the subjects and all the hand gestures, we found that, in the EMG-ACC sensor fusion system, EMG only contributed 8.3% to the recognition results, and ACC contributed 91.7%. If only considering the wrist movement gesture, EMG contributed 7.8%, and ACC contributed 92.2%; while if only considering the finger gesture, the EMG contributed 8.8%, and the ACC contributed 91.2%. Each EMG sensor contributed 0.69% to the recognition result. Compared with the healthy groups (0.63%), EMG sensors on amputees were 9.5% more effective. In addition, in many cases, amputees only have motion intention but lack actual movements, making it hard for the ACC sensor to collect motion signals; therefore, EMG is more important to the amputees than to the healthy group.

Sensor placement location showed a great impact on the results. For amputees, in the EMG-ACC fusion system, location A contributed most to the recognition results (56.1%), location C contributed 30.2%, and sensors at location B contributed least to the recognition results (13.7%). Also, the gesture categories did not show any impact on the locations' contribution distribution. However, for a single sensor, the amputees showed a significant difference from the healthy group. For amputees, if a sensor was placed on location A, it contributed 27.8% to the final recognition results, and 24.0% on location B, but 51.2% on location C (the gesture categories did not show any impacts on the locations' contribution distribution).

As shown above, the sensors placed on the biceps brachii and triceps brachii are vital for amputees. Although the forearm of the amputees still exists, the lack of hand motions will lead to the atrophy of the forearm muscles. However, in order to fit the new living conditions, they will try to use the forearm to replace part of the hand functions, and during this process, the biceps brachii and triceps brachii on the affected arm will become stronger. Therefore, the biceps brachii and triceps brachii showed more importance to amputees than to the healthy groups in hand gesture recognition.

With the explanation results, we proposed an optimized EMG-ACC sensor fusion solution for amputees. In this solution, all the ACC sensors and the EMG sensors that were placed on the biceps brachii, and triceps brachii were kept. We tested the proposed solution with the same experimental protocol. The result showed that with the optimized sensor fusion system, the number of sensors in the fusion system was reduced by 40%, and the recognition accuracy was increased to 79.9%. Compared with the ACC-only solution (accuracy 76.8%) and EMG-ACC solution (accuracy 79.2%), the IG XAI method effectively simplified the sensor fusion system and reduced the cost of hardware.

D. Comparative Analysis and Future Work

Compared with previous studies, our work filled the gap of the explainable multimodal sensor fusion in HMI. Li et al. [56] proposed a two-channel region-based CNN for explainable vision-based hand gesture recognition, which is significantly different from wearable-based methods. Lee et al. [57] proposed an explainable deep learning model for EMG-based finger angle estimation using attention but did not include multimodal sensor fusion or hand gesture recognition. Gozzi et al. [7] and Gulati et al. [58] utilized XAI methods to explain EMG-based hand gesture recognition; however, they did not include multimodal sensor fusion or a quantification explanation validation. Therefore, to our best knowledge, this is the first study to focus on the explainability of multimodal sensor fusion in HMI applications.

The influence of the target model's structure on the explanation results still needs to be investigated. From our preliminary experiment, if the target model changes slightly (e.g., the number of convolutional layers, kernel size, the number of fully collected layers' units, and so on), the attribution results stay

almost the same. However, if the target models are specially designed for optimization of the original attribution of the input data, the previous results might significantly change. These models include the attention mechanism, adversarial neural network, and other algorithms that adjust the weights of different data channels. With the development of deep learning technology, the structure of the target model will also become more and more complex, bringing a huge challenge to a generalized explanation result. Future work should include building a model-independent XAI framework or performing a deeper analysis of the different model structures.

VI. Conclusion

This article is the first work that utilized XAI methods to explain the working principle of multimodal sensor fusion systems in hand gesture recognition. Four attribution algorithms and four quantitative evaluation algorithms were performed on data of 17 hand gestures from 60 healthy subjects and 11 amputees to explore the working mechanism behind the multimodal system. Based on our proposed methods, the target system's redundancy is significantly reduced by 40%. With the cross-validation between the XAI result and physiological evidence, the working principle of the sensor fusion system is also more transparent. In addition, our work tries to maintain high universality to ensure the reliability of the result, and we intend to further investigate the effect of different model structures. In the end, our work could provide urgently needed information to the community to help improve HMI systems by reducing complexity and revealing explainable information that is typically hidden within deep neural networks, thereby facilitating patients' daily use of prosthetic hands and helping improve their quality of life.

References

[1] X. Song et al., "Proposal of a wearable multimodal sensing-based serious games approach for hand movement training after stroke," *Frontiers Physiol.*, vol. 13, p. 1065, Jun. 2022.

[2] P. Kang, S. Jiang, and P. B. Shull, "Synthetic emg based on adversarial style transfer can effectively attack biometric-based personal identification models," *bioRxiv*, 2022.

[3] S. Jiang, P. Kang, X. Song, B. Lo, and P. Shull, "Emerging wearable interfaces and algorithms for hand gesture recognition: A survey," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 85–102, 2022.

[4] S. Jiang et al., "Feasibility of wrist-worn, real-time hand, and surface gesture recognition via sEMG and IMU sensing," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3376–3385, Aug. 2018.

[5] T. Fukuda, "Cyborg and bionic systems: Signposting the future," *Cyborg Bionic Syst.*, vol. 2020, pp. 1–2, Sep. 2020.

[6] H. Wang, P. Kang, Q. Gao, S. Jiang, and P. B. Shull, "A novel PPG-FMG-ACC wristband for hand gesture recognition," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5097–5108, Oct. 2022.

[7] N. Gozzi, L. Malandri, F. Mercorio, and A. Pedrocchi, "XAI for myo-controlled prosthesis: Explaining EMG data for hand gesture classification," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108053.

[8] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Inf. Fusion*, vol. 81, pp. 59–83, May 2022.

[9] A. Tocchetti and M. Brambilla, "The role of human knowledge in explainable AI," *Data*, vol. 7, no. 7, p. 93, Jul. 2022.

[10] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, Dec. 2016.

[11] P. Kang, S. Jiang, P. B. Shull, and B. Lo, "Feasibility validation on healthy adults of a novel active vibrational sensing based ankle band for ankle flexion angle estimation," *IEEE Open J. Eng. Med. Biol.*, vol. 2, pp. 314–319, 2021.

[12] G. Riccardo et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[13] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.

[14] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," 2021, *arXiv:2101.05307*.

[15] R. Ghaeini, X. Z. Fern, and P. Tadepalli, "Interpreting recurrent and attention-based neural models: A case study on natural language inference," 2018, *arXiv:1808.03894*.

[16] P. Kang, J. Li, B. Fan, S. Jiang, and P. B. Shull, "Wrist-worn hand gesture recognition while walking via transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 952–961, Mar. 2022.

[17] J. Li and Q. Wang, "Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 79, pp. 229–247, Mar. 2022.

[18] J. Li, P. Kang, T. Tan, and P. B. Shull, "Transfer learning improves accelerometer-based child activity recognition via subject-independent adult-domain adaption," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2086–2095, May 2022.

[19] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[20] S. Ortega-Avila, B. Rakova, S. Sadi, and P. Mistry, "Non-invasive optical detection of hand gestures," in *Proc. 6th Augmented Hum. Int. Conf.*, Mar. 2015, pp. 179–180.

[21] Y. Huang, X. Yang, Y. Li, D. Zhou, K. He, and H. Liu, "Ultrasound-based sensing models for finger motion classification," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1395–1405, Sep. 2018.

[22] S. Kanoga, A. Kanemura, and H. Asoh, "Are armband sEMG devices dense enough for long-term use?—Sensor placement shifts cause significant reduction in recognition accuracy," *Biomed. Signal Process. Control*, vol. 60, Jul. 2020, Art. no. 101981.

[23] S. Jiang, Q. Gao, H. Liu, and P. B. Shull, "A novel, co-located EMG-FMG-sensing wearable armband for hand gesture recognition," *Sens. Actuators A, Phys.*, vol. 301, Jan. 2020, Art. no. 111738.

[24] E. Ceolini et al., "Hand-gesture recognition based on EMG and event-based camera sensor fusion: A benchmark in neuromorphic computing," *Frontiers Neurosci.*, vol. 14, p. 637, Aug. 2020.

[25] A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements," *J. NeuroEng. Rehabil.*, vol. 14, no. 1, pp. 1–14, Dec. 2017.

[26] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[27] P. Kang, J. Li, S. Jiang, and P. B. Shull, "A visual variability and visuo–tactile coordination inspired child adaptation mechanism for wearable age group recognition and activity recognition," *Adv. Intell. Syst.*, Oct. 2022, Art. no. 2200236.

[28] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.

[29] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*.

[30] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 12136–12145.

[31] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 5634–5641.

[32] X. Li et al., "Adversarial multimodal representation learning for click-through rate prediction," in *Proc. The Web Conf. 2020*, vol. 2020, pp. 827–836.

[33] Z. Zhang et al., "Neural machine translation with universal visual representation," in *Int. Conf. Learn. Representations*, 2019.

[34] K. Dhamdhere, M. Sundararajan, and Q. Yan, "How important is a neuron?" 2018, *arXiv:1805.12233*.

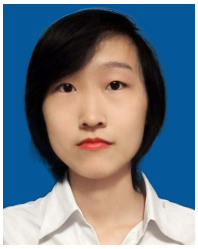
[35] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[37] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.
[38] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *J. Comput. Oper. Res.*, vol. 36, no. 5, pp. 1726–1730, 2009.
[39] A. Hedström et al., "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations," 2022, *arXiv:2202.06861*.
[40] M. Hashizume, "Perspective for future medicine: Multidisciplinary computational anatomy-based medicine with artificial intelligence," *Cyborg Bionic Syst.*, vol. 2021, pp. 1–3, Jan. 2021.
[41] D. S. Watson et al., "Clinical applications of machine learning algorithms: Beyond the black box," *Brit. Med. J.*, vol. 364, p. l886, Mar. 2019.
[42] P. V. Molle, M. D. Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, and B. Dhoedt, "Visualizing convolutional neural networks to improve decision support for skin lesion classification," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Cham, Switzerland: Springer, 2018, pp. 115–123.
[43] C. Biffi et al., "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 464–471.
[44] W. Jin, X. Li, and G. Hamarneh, "Evaluating explainable AI on a multimodal medical imaging task: Can existing algorithms fulfill clinical requirements?" in *Proc. Assoc. Advancement Artif. Intell. Conf. (AAAI)*, 2022, pp. 11945–11953.
[45] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[46] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*.
[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[48] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," 2020, *arXiv:2005.00631*.
[49] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10967–10978.
[50] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise explanations of neural networks using adversarial training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1383–1391.
[51] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 9505–9515.
[52] M. Atzori et al., "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, no. 1, pp. 1–13, 2014.
[53] R. N. Khushaba and K. Nazarpour, "Decoding HD-EMG signals for myoelectric control–how small can the analysis window size be?" *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8569–8574, Oct. 2021.
[54] J. Chen, S. Bi, G. Zhang, and G. Cao, "High-density surface EMG-based gesture recognition using a 3D convolutional neural network," *Sensors*, vol. 20, no. 4, p. 1201, Feb. 2020.
[55] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, Jan. 1993.
[56] P. Li and Z. Lu, "A novel art gesture recognition model based on two channel region-based convolution neural network for explainable human-computer interaction understanding," *Comput. Sci. Inf. Syst.*, vol. 19, no. 3, pp. 1371–1388, 2022.
[57] H. Lee, D. Kim, and Y.-L. Park, "Explainable deep learning model for EMG-based finger angle estimation using attention," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1877–1886, 2022.
[58] 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.

Peiqi Kang (Student Member, IEEE) received the B.E. degree in mechanical engineering and automation from Chongqing University, Chongqing, China, in 2019, and the M.E. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022, where he is currently pursuing the Ph.D. degree in mechanical engineering.

His research interests include human–machine interaction, intelligent sensing, and robotics.



Jinxuan Li (Student Member, IEEE) received the B.S. degree in mechanical design and manufacturing and its automation from Hunan University, Changsha, China, in 2019. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China.

Her research interests include wearable sensors, human activity recognition, and joint angle estimation.

Shuo Jiang (Member, IEEE) received the B.E. degree (Hons.) in mechatronic engineering from Zhejiang University, Hangzhou, China, in 2015, and the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2020.

From September 2019 to September 2020, he was a Visiting Scholar with Imperial College London, London, U.K. He is currently an Assistant Professor with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai. His research interests include human–machine interaction, intelligent sensing, and robotics.

Peter B. Shull (Member, IEEE) received the B.S. degree in mechanical engineering and computer engineering from LeTourneau University, Longview, TX, USA, in 2005, and the M.S. and Ph.D. degrees in mechanical engineering from Stanford University, Stanford, CA, USA, in 2008 and 2012, respectively.

From 2012 to 2013, he was a Post-Doctoral Fellow with the Bioengineering Department, Stanford University. He is currently a Professor of mechanical engineering with Shanghai Jiao Tong University, Shanghai, China. He has performed pioneering research involving human–computer interaction, hand gesture recognition, wearable systems, and real-time movement sensing and feedback to improve human health and performance in medical and sports applications. He has 18 competitive research grants, authored 86 peer-reviewed journal articles and conference papers, and delivered 55 academic technical presentations in English and Chinese. He has been the Primary Academic Advisor for 26 master's, doctoral, and Post-Doctoral Researchers.

Dr. Shull is an Associate Editor for *npj Digital Medicine* (Nature), the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, the IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, and *Wearable Technologies*.