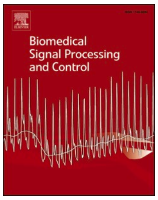
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



ViT-LLMR: Vision Transformer-based lower limb motion recognition from fusion signals of MMG and IMU

Hanyang Zhang [a], Ke Yang [a], Gangsheng Cao [a], Chunming Xia [a,b,*]

[a] *Department of Mechanical Engineering, East China University of Science and Technology, Shanghai 200237, China*
[b] *School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China*

ARTICLE INFO

Keywords:
Mechanomyography
Vision Transformer
Attention mechanism
Signal fusion

ABSTRACT

One of the key problems in lower limb-based human–computer interaction (HCI) technology is to use wearable devices to recognize the wearer's lower limb motions. The information commonly used to discriminate human motion mainly includes biological and kinematic signals. Considering that unimodal signals do not provide enough information to recognize lower limb movements, in this paper, we proposed a Vision Transformer (ViT)-based architecture for lower limb motion recognition from multichannel Mechanomyography (MMG) signals and kinematic data. Firstly, we applied the self-attention mechanism to enhance each input channel signal. Then the data was fed into ViT model. Vision Transformer-based Lower Limb Motion Recognition (ViT - LLMR) architecture proposed in this paper can avoid the model training problems such as autonomous feature extraction and feature selection for machine learning, and the model can recognize eight lower limb motions containing six subjects with an accuracy of 94.62%. In addition, we analyzed the generalization ability of the model when undersampling and only collecting fragment signals. In conclusion, the proposed ViT - LLMR architecture could provide a basis for practical applications in different HCI fields.

1. Introduction

Human lower limb motion pattern recognition technology has been applied to multiple fields related to human–computer interaction (HCI) such as medical monitoring, auxiliary rehabilitation training, intelligent prosthetics, and exoskeleton robots. In recent years, with the development of artificial intelligence, great progress has been made in the study of human lower limb motion pattern recognition. Due to the robustness to the environment, some physiological information related to the movement process gradually replaces the image to complete the motion classification, such as Electromyography (EMG) [1,2], Mechanomyography (MMG) [3], and some kinematic signals [4–6].

EMG is a signal generated by neuromuscular excitation and bioelectrical release during voluntary movement of the human body, which is commonly used in wearable devices for lower limb motion recognition. EMG has been widely used in describing both neuromuscular activities and muscular morphology [7]. The EMG signal on the skin surface is called the surface electromyography (sEMG) signal. Since sEMG is non-invasive, it has become an ideal signal source in the field of HCI. sEMG signals have been used for motion recognition of single joints of lower limbs: ankle joint [8,9], knee joint [10], and hip joint [11], as well as complex lower limb motion recognition involving multiple joints [12].As a counterpart of sEMG signal, MMG is a low-frequency mechanical signal generated by lateral vibration during muscle movement [13]. MMG signal has a high signal-to-noise ratio and it is immune to changes in skin impedance [14–16], and complex interactions of mechanical signals within the arm can produce repeatable patterns [17]. MMG is also widely used in disease diagnosis [18], muscle strength estimation [19], and motion pattern recognition [3]. At present, the studies on motion recognition based on MMG signals focus on the upper limbs [20], and there are few types of research on lower limbs.

In addition to biological signals, some kinematic signals collected by inertial measurement units (IMUs) [4,5], micro-triaxial flow sensors [5], gyroscopes [6], and force sensors [6] are also used to identify different movements of human. Using kinematic signals can obtain more stable recognition results for different movements of lower limbs, but motion segment detection and start prediction are difficult to achieve. In addition, the recognition accuracy using only kinematic signals is not ideal for motions with similar movements [20].

The research validated that the fusion of multiple modal signals is conducive to the improvement of motion recognition accuracy. Khomami et al [21] extracted features from sEMG and IMU respectively, and

[*] Corresponding author.

https://doi.org/10.1016/j.bspc.2022.104508
Received 3 September 2022; Received in revised form 10 November 2022; Accepted 18 December 2022
Available online 5 January 2023

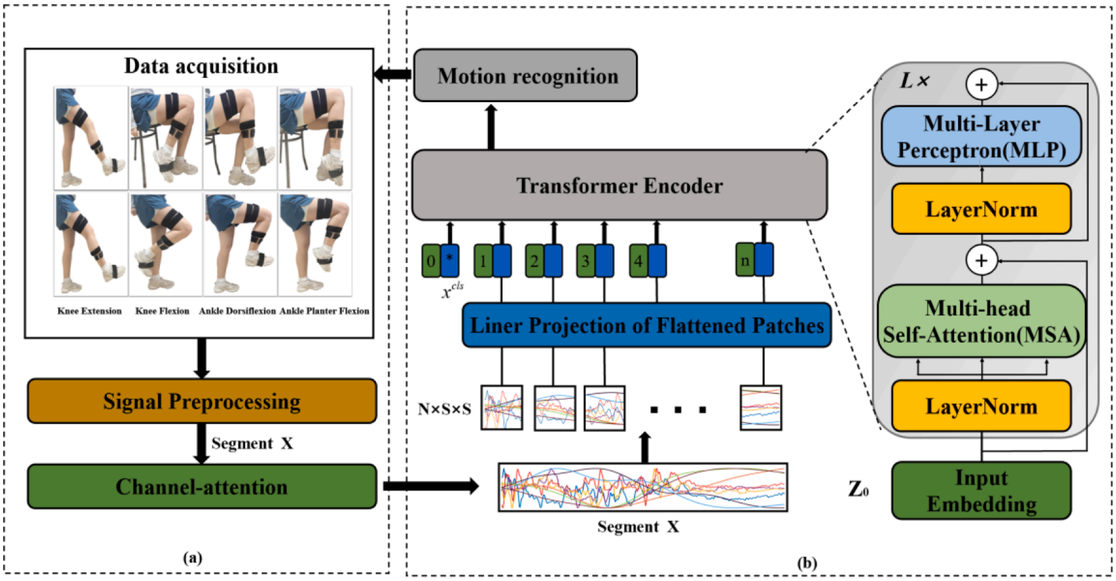


Fig. 1. The proposed ViT-LLMR architecture.

then used K nearest neighbors (KNN) to classify 20 sign language actions with an average accuracy of 96.13%. Zhou et al [22] performed dimensionality reduction on the extracted features of sEMG and IMU through principal component analysis (PCA) to form the fused feature set, and then achieves the classification of five lower limb movements by a support vector machine (SVM) classifier. Ai et al [23] extracted the time domain features and wavelet coefficients of sEMG and used dynamic time programming (DTW) distance for feature extraction of acceleration signals. Linear discriminant analysis (LDA) and SVM were used to classify five lower limb movements respectively. These results showed that using fused features in motion recognition may achieve better results than using sEMG signals or accelerometer signals only.

The current process of recognizing motion based on the fusion of different modal signals includes: noise filtering, signal segmentation, feature extraction and selection from different modal signals respectively, and motion classification by machine learning (ML) algorithms. ML algorithms commonly used for lower limb motion pattern recognition include LDA [23], SVM [23], KNN [24], and decision tree (DT) [25]. ML classifiers have some limitations: finding the optimal set of features is a very time-consuming task that requires expertise [26], and their performance degrades when applied to large-scale datasets.

In recent years, deep learning (DL) methods have become useful tools for motion pattern recognition, unlike ML algorithms that need to manually extract expert-defined features from the input data for classification, DL can automatically extract high-level abstract features from the input data while using multiple hidden layers. And DL has shown good performance on large datasets [27]. Convolutional neural network (CNN) is the most widely used DL structures for motion recognition based on biomedical signals. CNN performs much better than traditional methods (KNN, SVM, and LDA) in EMG, MMG, or IMU-based classification [28–31]. Modified CNN models are also applied to action recognition based on fused signals. Xu et al [32] used matrix counting method and time window amplitude method to convert sEMG and IMU into images then used dual-stream CNN for feature extraction, fusion, and classification of surface EMG signals and IMU images, respectively. The experimental results showed that the average recognition accuracy of the method was 95.78% for the six gestures of five subjects. Kwon et al [33] input the sEMG and IMU signal of forearm to two independent CNN networks respectively, and then determined the motion type based on the output results. However, CNN only covers the spatial domain of the signal and ignores the sequential nature [34], therefore, signals need to be transformed into images when they are input to CNN and the accuracy of the classification is highly dependent on the quality of the images, while the best procedure for generating images from 1D biosignals such as MMG/EMG is unknown [35]. To solve this problem, recurrent neural network (RNN) is proposed as a neural network for processing sequential data. RNN remembers the previous information. In the field of motion recognition, the deformed structure of RNN, Long Short-Term Memory (LSTM) neural network, has been widely used for EMG and IMU based motion recognition [36,37]. However, due to the sequential nature of LSTMs, parallelization is not supported in the training phase, causing a long training process.

Transformer was proposed in 2017 [38], it has excellent performance in natural language processing (NLP) [39], computer vision (CV) [40], speech processing [41], and anomaly detection [42]. In the field of biosignals, Song et al [43] first proposed a transformer-based electroencephalography (EEG) decoding architecture, which mainly relies on attentional mechanisms to learn the spatial and temporal characteristics of EEG signals. Rahimian et al [44] proposed a Transformer-based architecture to recognize upper-limb hand gestures from sEMG. In the last few years, many variants of the Transformer have been proposed to significantly improve the state-of-the-art performance for various tasks. In this paper, we designed a Vision Transformer-based architecture [40] to perform lower limb motion recognition from the fusion signals of MMG and IMU.

The main contributions of this paper can be summarized as follows:

A Vision Transformer-based architecture, called the ViT -LLMR architecture, was first designed to fuse multichannel MMG signals and IMU data for the recognition of lower limb motions.

Based on the self-attentive mechanism, the channel importance of different channel signals in different modes is weighted, which improves the defects of previous methods that ignore the importance of different channels. This method can be used for pre-processing before the fusion of different modal signals.

The effects of signal sampling frequency and signal length on the classification accuracy in the ViT-LLMR architecture were analyzed to provide a research basis for the practical application of the model.

The rest of this paper is organized as follows. Section 2 describes the overall architecture, including the data acquisition equipment, experimental procedure, data preprocessing, and the details of ViT-LLMR. We present the dataset, experiment details, and compare the results in Section 3. A detailed discussion is presented in Section 4, and finally, conclusions are drawn in Section 5.

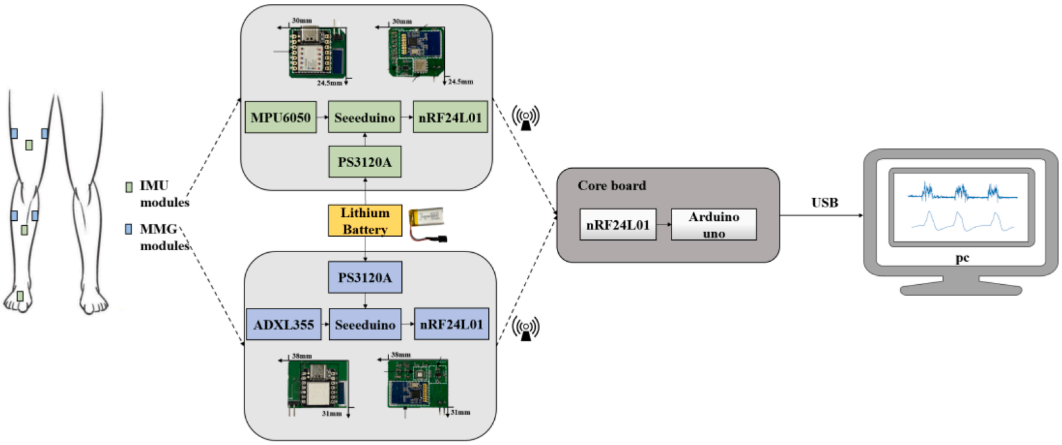


Fig. 2. The block diagram and prototypes of the wireless data acquisition system for MMG and IMU.

2. Method

In this paper, we proposed Vision Transformer (ViT) to fuse MMG signals and IMU data for lower limb motion recognition. The overall architecture is shown in Fig. 1 and consists of four parts. First, MMG and IMU data of the lower limb were collected synchronously using self-developed wireless devices, and then the MMG and IMU data were preprocessed, including noise filtering and signal segmentation; then different channels were weighted using channel attention so that the model can focus on the more relevant channels and ignore the irrelevant ones. Finally, a ViT-based architecture was designed to achieve lower limb motion recognition of the fused signals.

2.1. Data acquisition

2.1.1. Hardware description

The wireless data acquisition system consists of MMG modules and IMU modules with an independent power supply. As shown in Fig. 2, the proposed system can achieve wireless data acquisition and transmission when a person performs any action. The number and type of acquisition modules can be adjusted according to the usage scenario. Each module is independent to avoid interference with each other. Depending on the muscle positions of different subjects, the modules can be placed in different positions of the elastic band, and then the elastic band can be wrapped around the corresponding positions of the subject's thighs and calves, or directly pasted on the muscles.

ADXL355 sensor (Analog Devices, MASS, USA) was selected to collect the MMG signal because of the advantages of high precision, low offset drift, and low power consumption. The size of microcontroller unit (Seeed Studio, Shenzhen, CHN) is 20 × 17.5 × 3.5 mm, which can be flexibly used in various scenarios, especially in wearable devices. The lithium battery was used to power the system with a dimension of 35 × 20 × 6 mm. PS3120A (PULAN Technology, Hong Kong, CHN) was used for amplifying voltage, and nRF24L01(Nordic, Stockholm, Sweden) was selected for the wireless data transmission.

Mpu6050 inertial sensor (InvenSense, Sunnyvale, USA) was selected as the kinematic information acquisition sensor. Other parts of the module were consistent with the MMG module. MPU6050 consists of a gyroscope and a 3-axis accelerometer, a temperature sensor, and a Digital Motion Processor (DMP) module. The original data is converted into quad data by DMP, and the Euler angle is calculated according to the quaternion array. The definitions of quaternion and Euler angles are as follows:

$$q = [w, x, y, x]^T \tag{1}$$

The quaternion needs to be normalized, and was normalized by the constraints of (2):

$$|q|^2 = w^2 + x^2 + y^2 + z^2 = 1 \tag{2}$$

The posture of any object in three-dimensional space can be represented by Euler angles:

$$\begin{pmatrix} \phi \\ \theta \\ \varphi \end{pmatrix} = \begin{pmatrix} \mathrm{atan2}(2(\omega x + yz), 1 - 2(x^2 + y^2)) \\ \arcsin 2(wy - zx) \\ \mathrm{atan2}(2(\omega x + yz), 1 - 2(y^2 + z^2)) \end{pmatrix} \tag{3}$$

where ϕ is the roll angle, θ is the pitch angle and φ is the yaw angle. The yaw angle was not considered in this paper, because it will be affected by the rotation of the human body. The estimated value will have a large variation due to the restriction of measurement accuracy and become meaningless after a period of time.

2.1.2. Data collection

Six healthy males (age: 24 ± 3 years) with no history of neuromuscular disease volunteered to participate in the experiment. They were fully informed of the experimental content and they signed the informed consent. They were requested not to exercise vigorously in the 24 h before the experiment. Four MMG modules were placed on the abdomen of the four muscles of the lower limbs to detect MMG signals: vastus lateralis (channel 1), vastus medialis (channel 2), the lateral gastrocnemius (channel 3), medial gastrocnemius (channel 4), where the three IMU modules were placed on the thigh, shank, and instep. Subjects were asked to perform eight lower limb motions of the right lower limb: Knee Extension, Knee Flexion, Ankle Dorsiflexion, and Ankle Planter Flexion in sitting and standing positions, as shown in Fig. 1(a). Each motion was performed 100 times repeatedly for 4 s individually. On the premise of ensuring the integrity of signal acquisition, to reduce the amount of data, the sampling frequency was initially set to 250 Hz.

2.2. Signal preprocessing

MMG signals are bandwidth signals with frequencies between 2 and 120 Hz [13]. Since different muscles in different parts of the body and different movements of the same muscle generate MMG signals of different frequencies. The frequency of the MMG signal from lower limb movements is mainly the low frequency part below 50 Hz [45]. Therefore, the signal is filtered with a 2–50 Hz Finite Impulse Response (FIR) bandpass filter. Since most human lower extremity activities are in low frequencies, we applied a 5 Hz low-pass filter to the IMU signal [21].

In this paper, the Z-score method is used for the normalization of MMG and IMU data. The expression is as follows:

$$yi = \frac{xi - \overline{x}}{s} (i = 1, 2, \ldots n) \tag{4}$$

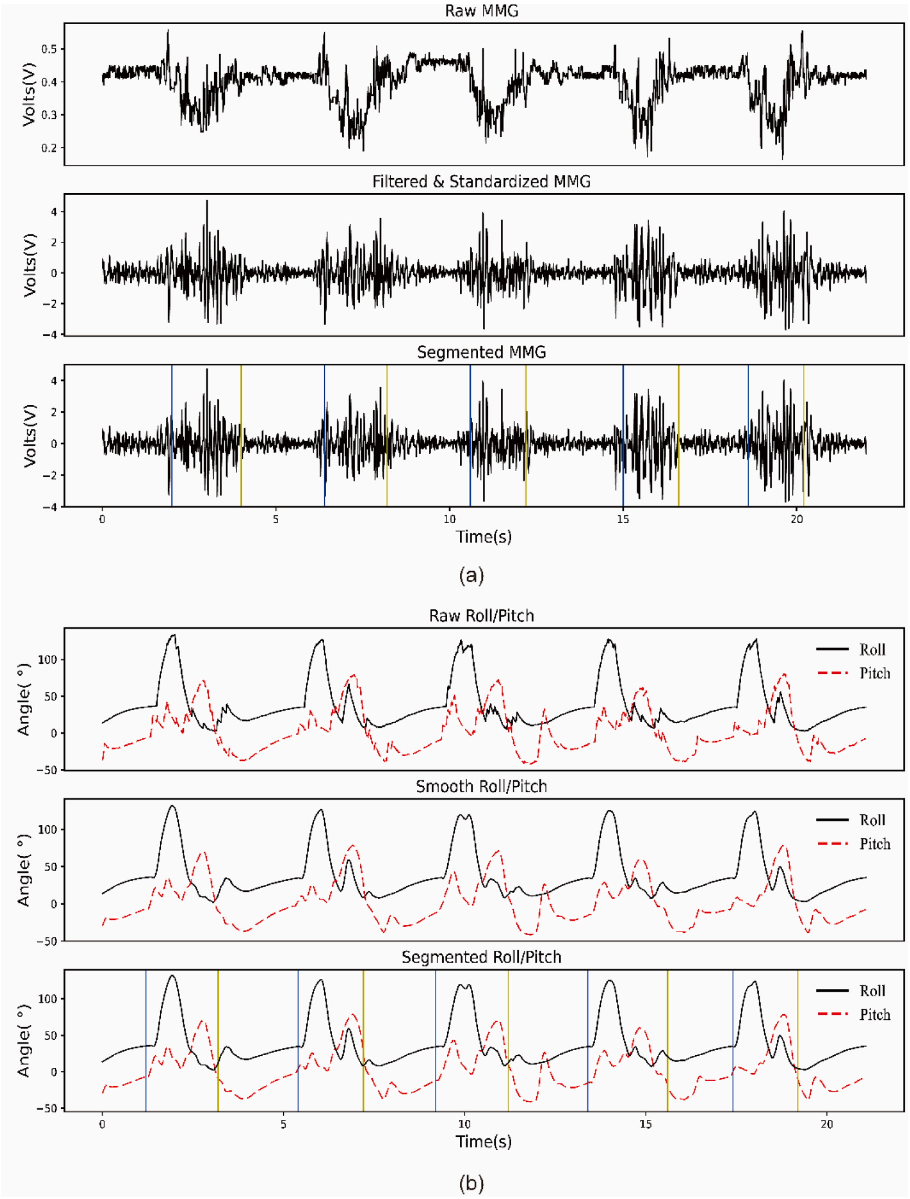


Fig. 3. The preprocessing and segmented of the signal of Knee Extension (a) MMG (b) IMU.

where $\overline{x}$ is the average value ofxi . s is the variance ofxi . The expression is as follows:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi-\overline{x})^{2}}(i = 1, 2, \dots n) \tag{5}$$

The acquired MMG data were divided into data sets by a sliding window of 200 ms using an average energy-based segmentation method as shown in Fig. 3(a). The average energy E for each set of data (all four channels) is compared to a predefined threshold TR. When $E(t1)$ and $E(t1+1)$ are larger than TR, but $E(t1-1)$ and $E(t1-2)$ are smaller than TR, the starting point $t1$ is determined. When $E(t2)$ and $E(t2+1)$ are smaller than TR, but $E(t2-1)$ and $E(t2-2)$ are larger than TR, the end point $t2$ is determined [46].

A differential threshold-based segmentation method was adopted on IMU signal, and the six-channel IMUs were divided into data groups by a sliding window of 200 ms as shown in Fig. 3(b). The difference value S for the six channels was calculated and compared with an empirically predefined threshold SR. When $S(t1)$ and $S(t1+1)$ are larger than SR, but $S(t1-1)$ and $S(t1-2)$ are smaller than SR, the starting point t1 is determined. The end point is determined when $S(t2)$ and $S(t2+1)$ are smaller than SR, but $S(t2-1)$ and $S(t2-2)$ are larger than SR [22].

The length of the segmented signal is between 1400 and 1800 ms. In order to facilitate subsequent research, the length of each motion frame was taken as 1600 ms.

The segmentation step transforms the MMG-MPU dataset into$D = \{\{X_i, y_i\}\}\ {}_{i=1}^{M}$, consisting of M segments, where the i^{th} segment is denoted by$Xi \in R^{S\times W}$, for $(1 \leq i \leq \mathrm{M})$, with its associated label denoted byyi. Here, S denotes the number of channels, and W shows the number of samples of each segment. In this paper, the value of S is 10, which contains 4 channels of MMG signals and 6 channels of IMU signals, and the value of W is 400, which is the number of samples per motion frame.

2.3. Channel attention

In previous biosignal-based studies, little attention has been paid to the importance of different input channels, this leads to mutual interference of information between channels, affecting recognition efficiency and accuracy. In this paper, the MMG and IMU signals of different

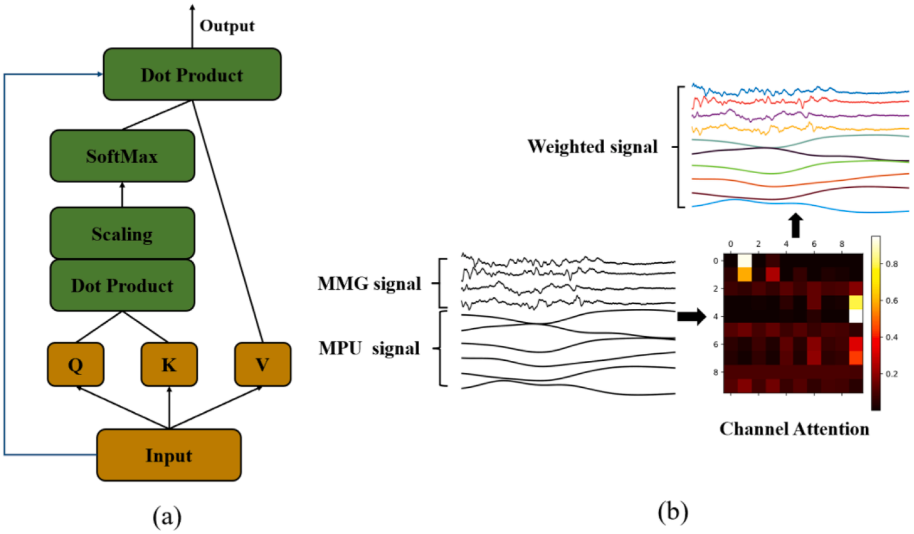


Fig. 4. The calculation process of channel attention(a) calculation principle(b) schematic diagram.

channels were weighted according to the channel attention method based on the scale dot product concern [38] proposed in [43] as shown in Fig. 4. We used the dot product to evaluate the correlation between channels. The input data X are first linearly transformed into vectors, queries (Q) and keys (K) and values (V), dh denotes the size of each vector in$Q$,K, and$V$. Q represents each channel that will be used to match with K represents all the other channels using the dot product. Then the result is scaled by $\sqrt{dh}$ and translated into the probabilities. The output weight score is assigned to V for the final representation using dot product. The whole process can be expressed as:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{dh}}\right)V \tag{6}$$

2.4. Vision transformer

2.4.1. Patch embeddings and position embeddings

We split the segmented input X into N non-overlapping patches. Since the number of channels is S, we set the size of each patch to ($S \times S$); thus, the number of patches will be equal to$N = W/S$. Each patch is then flattened into a vector$x_j^p \in R^{S^2}$, ($1 \leq \mathrm{j} \leq \mathrm{N}$). A linear projection is then applied to embed each vector into the model dimension d. For the linear projection, we used a matrix$E \in R^{S^2 \times d}$, which is shared among different patches. The output of this projection is called patch embeddings (Eq. (7) below). Similar to BERT's architecture [47], the beginning of the sequence of embedded patches is appended with a trainable [cls] tokenx^{cls}, to capture the meaning of the entire segmented input.

Both MMG and IMU signals are time series signals. If we change the time sequence, the meaning of the input signals may also change with it. And instead of processing the input in order, the transformer combines the information from the other elements by self-attention, so it is of great importance to encode the positions of the input time series into Transformers. A common design is to first encode positional information as vectors and then inject them into the model as an additional input together with the input time series. we add position embeddings denoted by $E^{pos} \in R^{(N+1) \times d}$ to the patch embeddings that will allow the transformer to capture the positional information. The formulation governing patch and position embeddings is given by:

$$Z0 = \left[x^{cls}; x_1^p E; x_2^p E; \ldots; x_N^p E\right] + E^{pos} \tag{7}$$

2.4.2. Transformer encoder

As shown in Fig. 1(b), the transformer encoder consists of L identical layers. L is the depth of the transformer. Each layer consists of two modules: a Multi-head Self-Attention (MSA) mechanism and a Multi-layer Perceptron (MLP) module. MSA is built based on the Self-Attention (SA) mechanism as shown in section 2.3. MSA enables the model to pay attention to information from different subspaces at different locations. The summation of weights is calculated as equation (8):

$$MSA(Q,K,V) = Concat(Head1, \cdots Headh)W^O \tag{8}$$

$$Headi = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), i = 1, 2 \cdots, h \tag{9}$$

In equation (8), h is defined as the number of heads, we applied the SA mechanism as equation (6) for each$Headi$, the calculation process is as equation (9). QW_i^Q, KW_i^K, VW_i^V indicates that the Multi-head Self-Attention mechanism use different weight matrices forQ, K, V.W^O indicates that the splicing results are linearly transformed to obtain the final multi-headed attention results.

The MLP module consists of two fully-connected linear layers and a Gaussian Error Linear Unit (GELU) activation function, to enhance the perception and non-linear learning capabilities of the model. LayerNorm (LN) is applied before every block.

$$Z^{'}l = MSA(LN(Zl-1)) + Zl - 1 \tag{10}$$

$$Zl = MLP(LN(Z^{'}l)) + Z^{'}l \tag{11}$$

$$y = LN(Z_L^0)\; l = 1 \ldots L \tag{12}$$

The final output of the transformer can be represented as follows:

$$ZL = [ZL0; ZL1; ZL2; \ldots; ZLN] \tag{13}$$

Finally, we apply a Linear Layer (LL) to$ZL0$,The number of output neurons is equal to the number of categories.

$$y = LL\left(LayerNorm(Z_L^0)\right) \tag{14}$$

3. Result

3.1. Dataset

Four channels of MMG signals and six channels of IMU signals were collected by the self-developed wireless acquisition system when six subjects performing eight lower limb motions. 100 sets per subject were acquired for each action. In total, 4800 sets of data were collected, 600

Table 1
Descriptions of ViT-LLMR architecture variants.

Model ID	Depth	Heads	Embedding size	Params	Accuracy (%)	STD (%)
1	1	8	16	6,362	90.89	1.31
2	2	8	16	9,642	92.70	0.49
3	3	8	16	12,922	92.90	0.59
4	**1**	**8**	**32**	**17,562**	**94.62**	**0.49**
5	2	8	32	30,656	94.87	0.81
6	3	8	32	43,360	94.28	0.63
7	1	8	64	158,752	94.80	1.03

sets for each motion, and 800 sets for each subject. In this paper, in order to investigate the generalization ability of the model, we focus on the analysis of 4800 sets of 6 subjects if not otherwise specified.

3.2. Experiment details

Our method was implemented with Python 3.8 and PyTorch library on a Geforce 3080Ti GPU. We evaluated different variants of the ViT-LLMR architecture. The details are summarized in Table 1. For all model variants, we set the size of the input patch to 10 × 10. All models were trained using Adam optimizer with betas = (0.9, 0.999), and the weight decay was set to 0.001. These models were trained with a batch size of 50. Cross-entropy loss was used for measuring classification performance. Ten-fold cross-validation was used to evaluate the final results.

3.3. Classification results

Table 1 shows the different ViT-LLMR architecture variants, we used the mean and standard deviation(std) of ten times ten-fold cross validation to evaluate the models' performance. It can be seen that at the same embedding size of 16, the accuracy of the models increases gradually as depth increases from 1 to 3, in the meantime the number of model trainable parameters has increased from 6362 to 12922. By increasing the embedding size from 16 to 32(Model 1 to Model 4), the accuracy of the models improved around 4% accompanied by a decrease in std, while the number of model parameters increased to 17562, which is more than it of model 3. Increasing the embedding size to 64 (model 7), the accuracy of the models improved only 0.18% with the increase of std, and the number of model parameters increased to 158752, which is around 10 times as many as model 4. In addition, we analyzed the model accuracy for different depths when embedding size is 32(Model 4 to Model 6). The accuracies of these three models fluctuate around 94.5%, while the model parameters rise from 17,562 to 43360.

As the increase of model parameters brings problems such as longer

Table 2
The result of Wilcoxon signed-rank test.

Model	2	3	4	5	6	7
1	0.00335*	0.00444*	0.00335*	0.00335*	0.00335*	0.00442*
2		0.65664	0.00328*	0.00444*	0.00444*	0.01637
3			0.00333*	0.00335*	0.00762	0.01637
4				0.42337	0.09116	0.32806
5					0.10951	0.04671*
6						0.78968

* Significant.

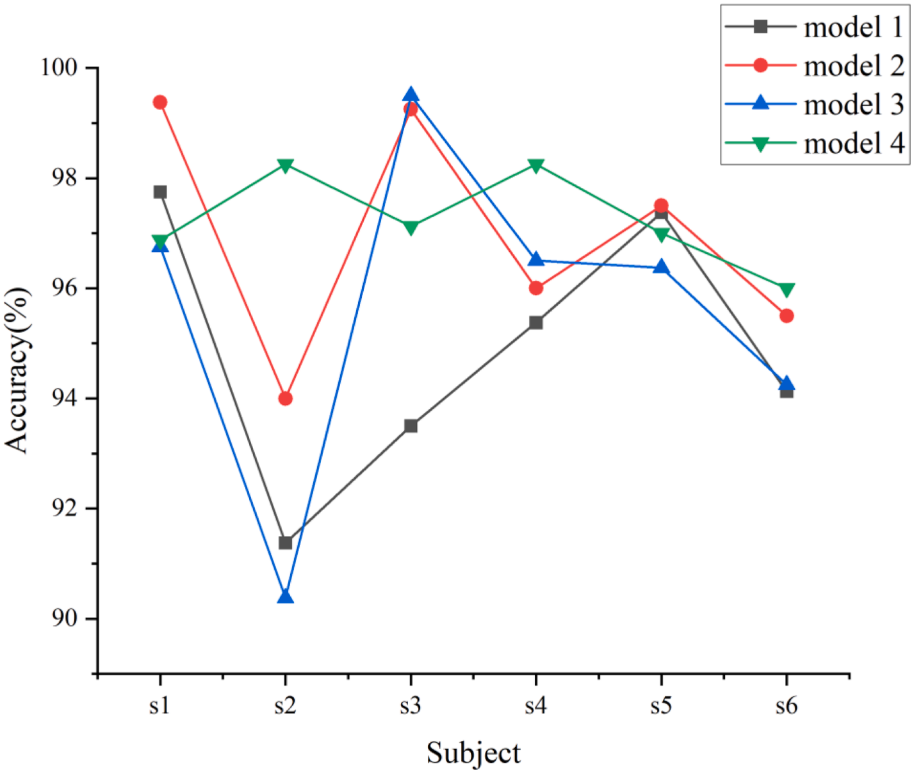


Fig. 5. Classification accuracies of different subjects of ViT-LLMR architecture variants.

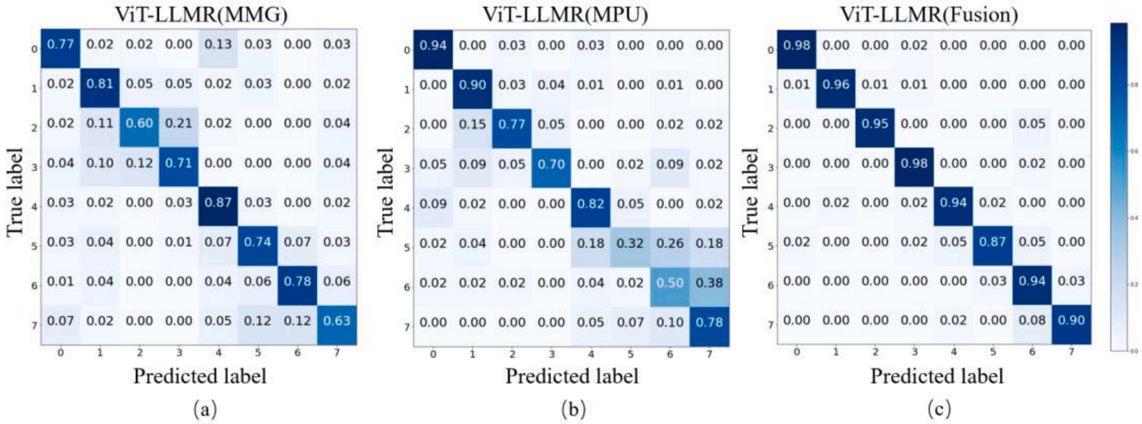


Fig. 6. Confusion Matrix for ViT-LLMR architecture (model 4) (a)MMG (b)IMU (c)Fusion signal.

Table 3
The Label of 8 Lower Limb Motions.

label	motion	
0	sitting	Knee Extension
1		Knee Flexion
2		Ankle Dorsiflexion
3		Ankle Plantar Flexion
4	standing	Knee Extension
5		Knee Flexion
6		Ankle Dorsiflexion
7		Ankle Plantar Flexion

Table 4
Comparison with representative methods.

Classifier		Accuracy (%)	STD (%)
ViT-LLMR	**Fusion signal**	**94.62**	**0.49**
	MMG signal	73.88	1.12
	MPU signal	71.63	4.25
	without channel attention	89.17	3.12
Traditional fusion method	LDA	70.70	2.12
	SVM	72.90	1.96
	KNN	69.87	3.19
	CNN [49]	88.01	0.78

training time, more memory required, and higher equipment requirements, we used Wilcoxon signed rank test [48] to show the significance level of different architecture variants as shown in Table 2. According to the results, the differences in accuracy between model 1 and other six models were considered to be statistically significant ($p < 0.005$), the difference in accuracy between model 2 and model 3 was considered not statistically significant ($p = 0.65664$, $p > 0.005$). The difference in accuracy between model 4 and models 5, 6, and 7 were considered not statistically significant ($p > 0.005$). When the depth reaches 3 and the embedding size reaches 32, the increase in depth and embedding size of architecture does not significantly improve the accuracy, while introduces more parameters, so in the following we only analyze model 1, 2, 3, and 4.

In addition, we analyzed the dataset of each subject separately, and the classification accuracy is shown in Fig. 5. It can be seen that the classification accuracy of model 4 is above 96% for all six subjects, and model 2 performs better than model 4 for subjects 1, 3, and 5, but performs poorly on subject 2, which is around 4% less accurate than model 4.

To analyze whether the fusion of MMG and IMU signals has a significant improvement on the classification accuracy, we compared the classification accuracy of the input as 4-channel MMG signals, 6-channel IMU data and fused 10-channel signals with the confusion matrix shown in Fig. 6. The labels and corresponding motions are shown in Table 3. It can be seen that if only based on the MMG signal, the classification accuracy for each motion is relatively average but not high, at around 70%, while based on the IMU data only, the classification accuracy is low for a few specific actions (Knee Flexion and Ankle Dorsiflexion in the standing state) and high for the rest motions. This may be because the changes of the thigh and shank positions of these two motions are similar, so it is difficult to judge only by the IMU data, while the MMG signal performs better because the muscle strength is clearly differentiated during the execution of these two motions. As can be seen in Fig. 6c, the fusion of the two signals resulted in a high recognition rate for all eight actions.

The classification accuracy of the input as 10-channel fusion signals, 4-channel MMG signals, 6-channel IMU data and 10-channel signals without channel attention were be compared. Then we compared ViT-LLMR on our dataset with the algorithms commonly used for lower limb motion recognition for comparison: the SVM [23], LDA [23], KNN [20,24], and CNN, and the CNN model refers to the structure proposed by [49]. As can be seen from Table 4, the classification accuracy of ViT-LLMR based on fusion signal is much higher than that of several other methods.

4. Discussion

In this study, we propose a transformer-based ViT-LLMR architecture that enables the recognition of lower limb movements by weighting the input channels using an attention mechanism and then feeding them into a ViT model. The model can directly use the original MMG and IMU signals without relying on manual feature extraction, which not only simplifies many complicated steps of signal preprocessing but also solves the problem of feature selection limitations. CNN is the most commonly used model in motion recognition based on biosignals, however, traditional CNN models aim to learn spatial features and cannot extract temporal features from time-series signals. To address this drawback, recent studies have proposed RNN, such as LSTM, to capture temporal information from MMG signals [50]. However, due to the sequential nature of RNN, it does not allow parallelization in the training phase, so the training speed is slow [44]. The transformer neural network architecture eliminates recurrence or convolution by using a self-attentive mechanism, it shows superior ability to deal with long-range dependencies [38].To verify the generalization ability of the model in practical applications, we also analyzed the classification accuracy in the case of signal undersampling. To follow the Nyquist’s rate, the sampling rate should be at least twice the highest frequency of the signal to avoid aliasing [51]. However, most of the current studies on MMG

Table 5
Classification accuracies for ViT-LLMR architectures variants. The STD represents the standard variation in accuracy over 10 times 10-fold cross-validation.

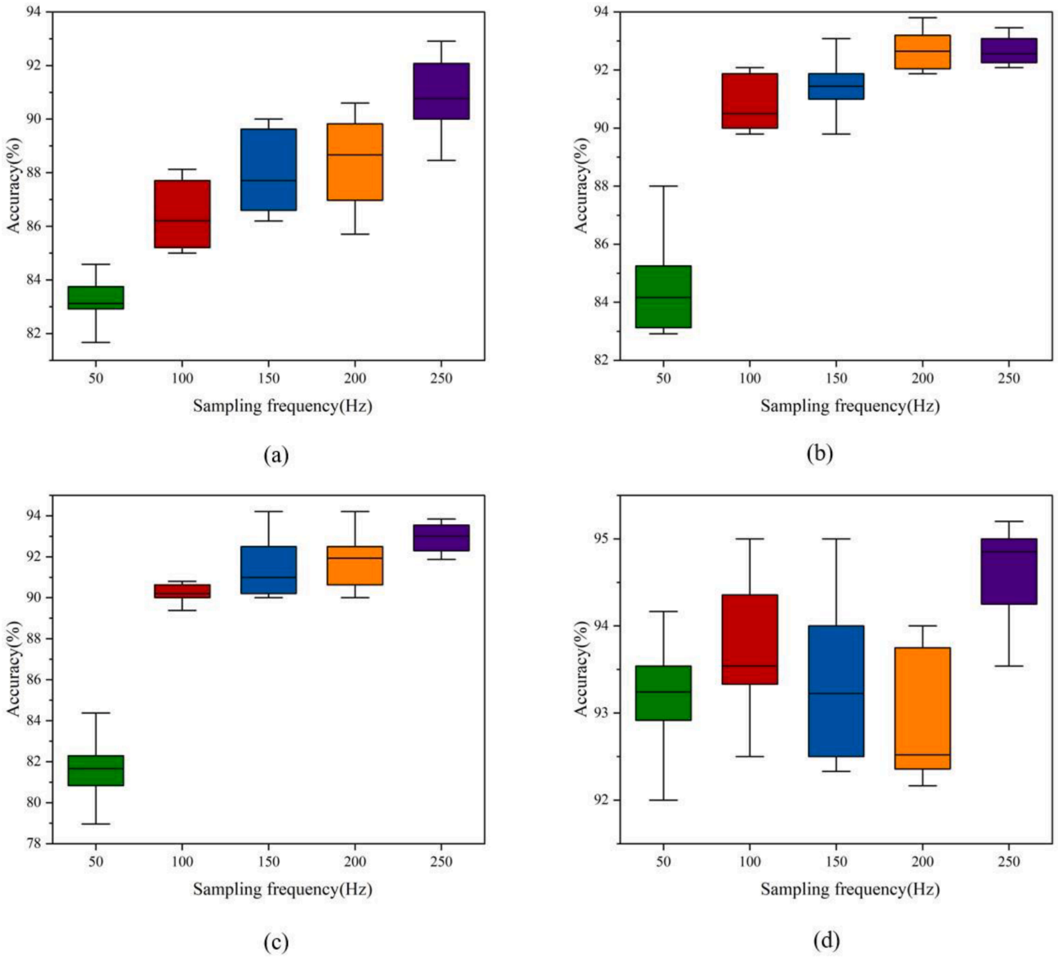
Sampling Frequency	Model ID	1	2	3	4
250 Hz	Accuracy (%)	90.89	92.70	92.90	94.62
	STD (%)	1.31	0.49	0.59	0.49
200 Hz	Accuracy (%)	88.40	92.77	91.88	92.86
	STD (%)	1.62	0.65	1.34	0.65
150 Hz	Accuracy (%)	87.90	91.07	91.48	93.31
	STD (%)	1.40	0.92	1.47	0.89
100 Hz	Accuracy (%)	86.37	90.77	90.33	93.65
	STD (%)	1.14	0.85	0.57	0.79
50 Hz	Accuracy (%)	83.27	84.43	81.60	92.96
	STD (%)	0.82	1.48	1.41	1.07

signals set the sampling frequency at 1 kHz [52,53], which is much higher than twice the frequency of MMG signals. A higher sampling rate means more data is collected per unit of time, and the requirements on the hardware equipment become more stringent. For example, the conversion (A/D) module, requires higher resolution, better dynamic characteristics, and stronger conversion performance. Correspondingly, the cost can be increased significantly [54], and the larger the amount of data, the lower the recognition speed of the model. To obtain the effect of signal sampling frequency on the classification accuracy of Vit-LLMR, we analyzed the classification accuracy of four models at 50 Hz, 100 Hz, 150 Hz, 200 Hz, and 250 Hz. The results are shown in Table 5 and Fig. 7.

It can be seen that in models 1–3, the accuracy is much lower than the accuracy of the remaining four at a sampling frequency of 50 Hz, while in model 4, they are less different. The lowest is also above 92%. In model 1,3,4, the average accuracy of classification at 250 Hz is the highest in all models, but in model 2, the average accuracy at 200 Hz is slightly higher than that at 250 Hz. This result shows that after choosing the appropriate model parameters, an appropriate reduction of the sampling frequency does not have a significant impact on the recognition efficiency of the model.

If the recognition of the action can be realized based on the part of the signal during the execution of the action, then the problem of delayed output results caused by obtaining the signal segment of the entire action can be solved. To obtain the effect of signal sampling length on classification accuracy, we analyzed the classification accuracy of four models at 100(0.4 s), 200(0.8 s), 300(1.2 s), 400(1.6 s). The schematic diagram of the signal fragment is shown in Fig. 8. Fig. 9 compares the classification accuracies for different sampling lengths. When only the data of the first 100 samples are input (0.4 s), the classification accuracy is low and is below 80%. When the number of input samples is 200 and 300, the classification accuracy increases and is above 80%. In particular, in models 2 and 4, the classification accuracy reaches more than 90% for the first 300 points of the input. The difference with the classification accuracy of the input complete signal is small (<2%). The results show that the recognition can be done successfully before the action is completed.

In this paper, we analyzed the accuracies of models developed by using data from all subjects and subject-specific models based on ten-fold cross-validation. One drawback of these two evaluation methods is that data from the same person will be used in both the training sets and test sets[55]**.** To better evaluate the generalization performance of the model, we trained the model with leave-one-subject-out (LOSO)



(a)

(b)

(c)

(d)

Fig. 7. The accuracy boxplots for all ViT-LLMR architecture variants at different sampling frequencies. (a)model1(b)model2(c)model 3(d)model 4.

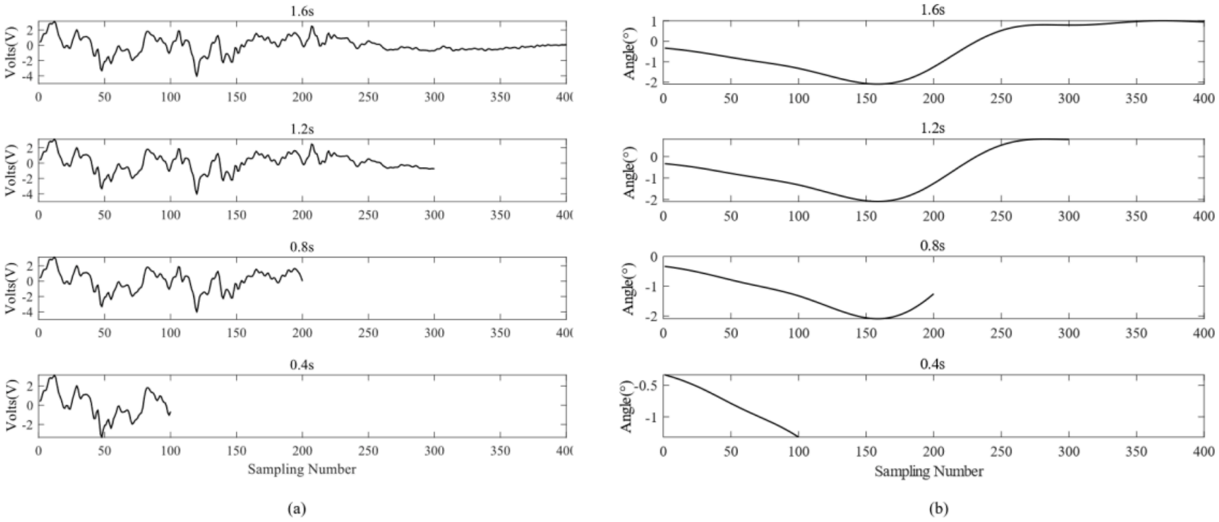


Fig. 8. The schematic diagram of the signal fragment(a)MMG(b)IMU.

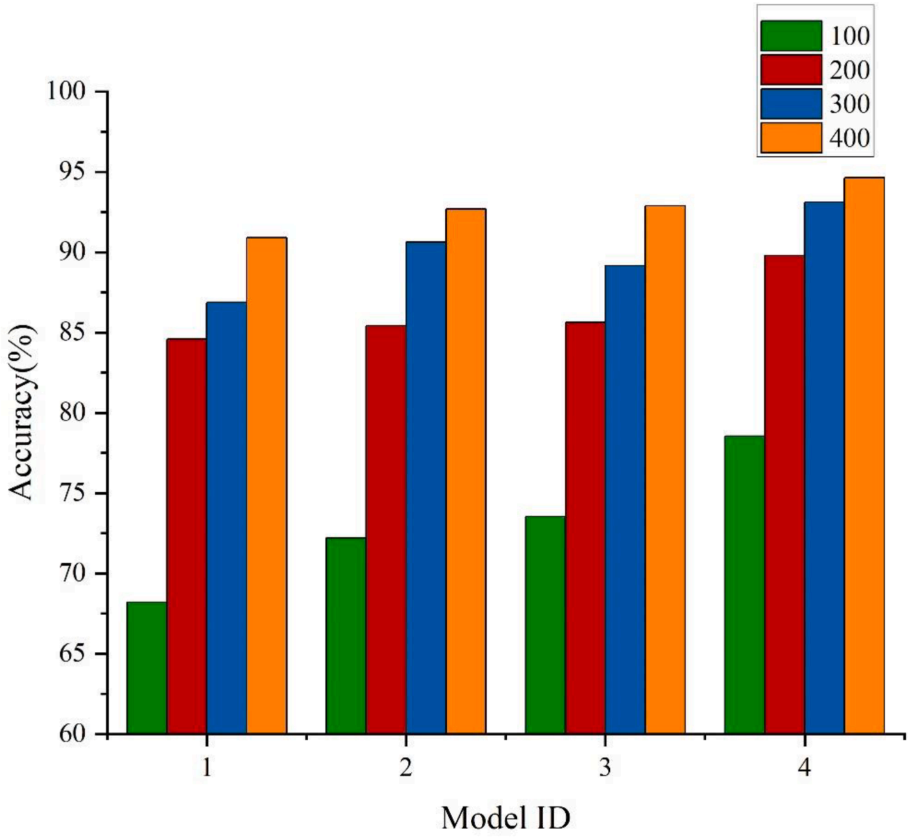


Fig. 9. The accuracy barplot for all ViT-LLMR architecture variants at different sampling lengths.

Table 6
Classification accuracies for ViT-LLMR architectures using leave-one-subject-out cross-validation (%).

Leave-out-Subject	1	2	3	4	5	6	Average
Model 1	75.88	64.25	71.23	65.38	76.25	74.38	71.23
Model 2	77.62	65.45	74.23	71.21	74.12	78.81	73.57
Model 3	76.21	66.57	76.78	73.89	79.12	82.22	75.80
Model 4	80.12	66.34	79.13	73.34	81.34	80.13	76.73

cross-validation**.** Table 6 shows that the accuracies of models using LOSO cross-validation were significantly lower than using 10-fold cross-validation due to the individual variability of each subject performing lower limb motions, for example, when subjects 2 and 4 were selected as the test set, the classification accuracies were lower. The average accuracies of the four models are between 70% and 80%.

In summary, the proposed ViT-LLMR model has better classification accuracy than traditional machine learning methods and CNN and maintains high recognition accuracy when undersampling and only use part of signals. The model has good generalization and robustness and can be better applied to human-computer interaction technologies:

exoskeleton control, virtual reality interaction, health detection, etc. In future research, we will also continue to investigate improving the generalization performance of the model when completely separating the source samples of the training set from the test set. We need to collect data of more motions, as only eight categories of movements commonly used in rehabilitation training have been studied so far. In addition, we will try to improve the model to enhance the classification accuracy under undersampling conditions.

5. Conclusion

We proposed a transformer-based ViT-LLMR architecture that enables the recognition of lower limb movements, and evaluate the generalization ability of ViT-LLMR based on a 10-fold cross-validation. We conducted a large number of experiments to analyze the effects of model parameters, signal sampling frequency, sampling length, and signal fusion on classification accuracy. The results show that when appropriate model parameters are chosen, appropriately decreasing the sampling frequency or reducing the length of the input signal has less effect on the classification accuracy, and also compare with the commonly used classification algorithms in this field, showing the superiority of ViT-LLMR.

CRediT authorship contribution statement

Hanyang Zhang: Conceptualization, Software, Visualization, Data curation, Writing – original draft. **Ke Yang:** Writing – review & editing, Methodology. **Gangsheng Cao:** Writing – review & editing. **Chunming Xia:** Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to express their gratitude to the volunteers participating in the experiment.

References

[1] F.K. Mukinda, et al., The crowded space of local accountability for maternal, newborn and child health: a case study of the South African health system, Health Policy Plan. 35 (3) (2020) 279–290.

[2] Y. Tao, et al., Multi-channel sEMG based human lower limb motion intention recognition method. in 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2019.

[3] A. Wołczowski, R. Zdunek, Electromyography and mechanomyography signal recognition: Experimental analysis using multi-way array decomposition methods, Biocybernetics Biomed. Eng. 37 (1) (2017) 103–113.

[4] M.M. Hamdi, et al. Lower limb motion tracking using IMU sensor network, in: 2014 Cairo International Biomedical Engineering Conference (CIBEC). IEEE, 2014.

[5] S. Liu, et al., A wearable motion capture device able to detect dynamic motion of human limbs, Nat. Commun. 11 (1) (2020) 1–12.

[6] J. Song, et al., Human body mixed motion pattern recognition method based on multi-source feature parameter fusion, Sensors 20 (2) (2020) 537.

[7] M. Reaz, M. Hussain, F. Mohd-Yasin, Techniques of EMG signal analysis: detection, processing, classification and applications (Correction), Biological procedures online 8 (1) (2006).

[8] D.C. Toledo-Pérez, et al., A study of movement classification of the lower limb based on up to 4-EMG channels, Electronics 8 (3) (2019) 259.

[9] M.S. Al-Quraishi, et al., Classification of ankle joint movements based on surface electromyography signals for rehabilitation robot applications, Med. Biol. Eng. Compu. 55 (5) (2017) 747–758.

[10] Y. Zhang, et al., Extracting time-frequency feature of single-channel vastus medialis EMG signals for knee exercise pattern recognition, PLoS One 12 (7) (2017) e0180526.

[11] Q. Li, Y. Song, Z. Hou, Estimation of lower limb periodic motions from sEMG using least squares support vector regression, Neural Process. Lett. 41 (3) (2015) 371–388.

[12] C. Tapia, O. Daud, J. Ruiz-del-Solar, EMG signal filtering based on independent component analysis and empirical mode decomposition for estimation of motor activation patterns, J. Med. Biol. Eng. 37 (1) (2017) 140–155.

[13] T.J. Herda, M.A. Cooper, Muscle-related differences in mechanomyography frequency–force relationships are model dependent, Med. Biol. Eng. Compu. 53 (8) (2015) 689–697.

[14] H.-B. Xie, Y.-P. Zheng, J.-Y. Guo, Classification of the mechanomyogram signal using a wavelet packet transform and singular value decomposition for multifunction prosthesis control, Physiol. Meas. 30 (5) (2009) 441.

[15] D.T. Barry, S.R. Geiringer, R.D. Ball, Acoustic myography: a noninvasive monitor of motor unit fatigue, Muscle Nerve: Off. J. American Association of Electrodiagnostic Medicine 8 (3) (1985) 189–194.

[16] M. Stokes, Acoustic myography: applications and considerations in measuring muscle performance, Isokinet. Exerc. Sci. 3 (1) (1993) 4–15.

[17] S. Wilson, R. Vaidyanathan, Gesture recognition through classification of acoustic muscle sensing for prosthetic control. Conference on Biomimetic and Biohybrid Systems, Springer, 2017.

[18] S.W. Jun, et al., Brief report: Preliminary study on evaluation of spasticity in patients with brain lesions using mechanomyography, Clin. Biomech. 54 (2018) 16–21.

[19] M. Jo, et al., Mechanomyography for the measurement of muscle fatigue caused by repeated functional electrical stimulation, Int. J. Precis. Eng. Manuf. 19 (9) (2018) 1405–1410.

[20] M.-K. Liu, et al., Hand gesture recognition by a MMG-based wearable device, IEEE Sens. J. 20 (24) (2020) 14703–14712.

[21] S.A. Khomami, S. Shamekhi, Persian sign language recognition using IMU and surface EMG sensors, Measurement 168 (2021), 108471.

[22] B. Zhou, et al., Accurate recognition of lower limb ambulation mode based on surface electromyography and motion data using machine learning, Comput. Methods Programs Biomed. 193 (2020), 105486.

[23] Q. Ai, et al., Research on lower limb motion recognition based on fusion of sEMG and accelerometer signals, Symmetry 9 (8) (2017) 147.

[24] A.F. Laudanski, S.M. Acker, Classification of high knee flexion postures using EMG signals, Work 68 (3) (2021) 701–709.

[25] D. Rodriguez, A. Piryatinska, X. Zhang, A neural decision forest scheme with application to EMG gesture classification. in 2016 SAI Computing Conference (SAI). IEEE, 2016.

[26] D. Xiong, et al., Deep learning for EMG-based human-machine interaction: A review, IEEE/CAA J. Autom. Sin. 8 (3) (2021) 512–533.

[27] K.-H. Park, S.-W. Lee, Movement intention decoding based on deep learning for multiuser myoelectric interfaces. in 2016 4th international winter conference on brain-computer Interface (BCI). IEEE, 2016.

[28] E.A. Chung, M.E. Benalcázar, Real-time hand gesture recognition model using deep learning techniques and EMG signals, in: 2019 27th European Signal Processing Conference *(EUSIPCO)*. IEEE, 2019.

[29] A. Waris, et al., The effect of time on EMG classification of hand motions in able-bodied and transradial amputees, J. Electromyogr. Kinesiol. 40 (2018) 72–80.

[30] F. Höflinger, et al., A wireless micro inertial measurement unit (IMU), IEEE Trans. Instrum. Meas. 62 (9) (2013) 2583–2595.

[31] H. Wu, et al., A CNN-SVM combined model for pattern recognition of knee motion using mechanomyography signals, J. Electromyogr. Kinesiol. 42 (2018) 136–142.

[32] L. Xu, et al., Gesture recognition using dual-stream CNN based on fusion of semg energy kernel phase portrait and IMU amplitude image, Biomed. Signal Process. Control 73 (2022), 103364.

[33] Y.D. Kwon, et al., Myokey: Surface electromyography and inertial motion sensing-based text entry in ar, in: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 2020.

[34] Y. Hu, et al., A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition, PLoS One 13 (10) (2018) e0206049.

[35] P. Tsinganos, et al. A Hilbert curve based representation of sEMG signals for gesture recognition. in 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, 2019.

[36] M. Simão, P. Neto, O. Gibaru, EMG-based online classification of gestures with recurrent neural networks, Pattern Recogn. Lett. 128 (2019) 45–51.

[37] O. Steven Eyobu, D.S. Han, Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network, Sensors 18 (9) (2018) 2892.

[38] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).

[39] J. Devlin, et al., Pre-training of deep bidirectional transformers for language understanding In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics, 2019, pp. 4171-86.

[40] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[41] L. Dong, S. Xu, B. Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. IEEE.

[42] Z. Chen, et al., Learning graph structures with transformer for multivariate time series anomaly detection in iot, IEEE Internet Things J. (2021).

[43] Y. Song, et al., Transformer-based spatial-temporal feature learning for eeg decoding. arXiv preprint arXiv:2106.11170, 2021.

[44] E. Rahimian, et al., TEMGNet: Deep Transformer-based Decoding of Upperlimb sEMG for Hand Gestures Recognition. arXiv preprint arXiv:2109.12379, 2021.

[45] J. Yu, Y. Zhang, C. Xia, Study of gait pattern recognition based on fusion of mechanomyography and attitude angle signal, J Mech Med Biol 20 (02) (2020) 1950085.

[46] W. Guo, et al., Mechanomyography assisted myoeletric sensing for upper-extremity prostheses: A hybrid approach, IEEE Sens. J. 17 (10) (2017) 3100–3108.

[47] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[48] E. Rahimian, et al., Fs-hgr: Few-shot learning for hand gesture recognition via electromyography, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 1004–1015.

[49] B.-Y. Su, et al., A CNN-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (5) (2019) 1032–1042.

[50] C. Xie, et al., A long short-term memory neural network model for knee joint acceleration estimation using mechanomyography signals, Int. J. Adv. Rob. Syst. 17 (6) (2020) 4463–14411.

[51] Nyquist and H., Certain Topics in Telegraph Transmission Theory. Proceedings of the IEEE, 90(2) (1928) 280-305.

[52] C.S.M. Castillo, et al., Wearable MMG-plus-one armband: Evaluation of normal force on mechanomyography (MMG) to enhance human-machine interfacing, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2020) 196–205.

[53] M.M. Ismail, et al., Hand motion pattern recognition analysis of forearm muscle using MMG signals, Bull Electrical Eng Informatics 8 (2) (2019) 533–540.

[54] H. Chen, et al., Exploring the relation between EMG sampling frequency and hand motion recognition accuracy. in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2017.

[55] D. Gholamiangonabadi, N. Kiselov, K. Grolinger, Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection, IEEE Access 8 (2020) 133982–133994.