applied sciences

MDPI

Article

A Hybrid Human Activity Recognition Method Using an MLP Neural Network and Euler Angle Extraction Based on IMU Sensors

Yaxin Mao [1], Lamei Yan [2], Hongyu Guo [1], Yujie Hong [1], Xiaocheng Huang [1,3,*] and Youwei Yuan [1,3]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China; 21051303@hdu.edu.cn (Y.M.); 21322108@hdu.edu.cn (H.G.); 21051613@hdu.edu.cn (Y.H.); yyw@hdu.edu.cn (Y.Y.)
[2] School of Media and Design, Hangzhou Dianzi University, Hangzhou 310018, China; ylm@hdu.edu.cn
[3] Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou 310018, China
* Correspondence: 221050090@hdu.edu.cn

Abstract: Inertial measurement unit (IMU) technology has gained popularity in human activity recognition (HAR) due to its ability to identify human activity by measuring acceleration, angular velocity, and magnetic flux in key body areas like the wrist and knee. It has propelled the extensive application of HAR across various domains. In the healthcare sector, HAR finds utility in monitoring and assessing movements during rehabilitation processes, while in the sports science field, it contributes to enhancing training outcomes and preventing exercise-related injuries. However, traditional sensor fusion algorithms often require intricate mathematical and statistical processing, resulting in higher algorithmic complexity. Additionally, in dynamic environments, sensor states may undergo changes, posing challenges for real-time adjustments within conventional fusion algorithms to cater to the requirements of prolonged observations. To address these limitations, we propose a novel hybrid human pose recognition method based on IMU sensors. The proposed method initially calculates Euler angles and subsequently refines them using magnetometer and gyroscope data to obtain the accurate attitude angle. Furthermore, the application of FFT (Fast Fourier Transform) feature extraction facilitates the transition of the signal from its time-based representation to its frequency-based representation, enhancing the practical significance of the data. To optimize feature fusion and information exchange, a group attention module is introduced, leveraging the capabilities of a Multi-Layer Perceptron which is called the Feature Fusion Enrichment Multi-Layer Perceptron (GAM-MLP) to effectively combine features and generate precise classification results. Experimental results demonstrated the superior performance of the proposed method, achieving an impressive accuracy rate of 96.13% across 19 different human pose recognition tasks. The proposed hybrid human pose recognition method is capable of meeting the demands of real-world motion monitoring and health assessment.

Keywords: human activity recognition; HAR system; IMU sensors; FFT; MLP neural network

Citation: Mao, Y.; Yan, L.; Guo, H.; Hong, Y.; Huang, X.; Yuan, Y. A Hybrid Human Activity Recognition Method Using an MLP Neural Network and Euler Angle Extraction Based on IMU Sensors. *Appl. Sci.* **2023**, *13*, 10529. https://doi.org/10.3390/app131810529

Academic Editors: Marc Kurz, Erik Sonnleitner and Clemens Holzmann

Received: 15 August 2023
Revised: 19 September 2023
Accepted: 20 September 2023
Published: 21 September 2023

1. Introduction

With the continuous advancement of technology and the increasing focus on physical well-being, sensor-based human activity recognition (HAR) has gained significant attention and has been widely applied. Human activity refers to the motion status of the human body in space, encompassing its position, angles, and trajectories. The recognition of human activity holds great potential across diverse domains, including medical rehabilitation [1], sports science, human–computer interactions, and gaming. In these fields, the attainment of precise and real-time human activity recognition plays a pivotal role in attaining predefined objectives.

The practical uses of traditional camera-based activity recognition systems are severely constrained by their low recognition accuracy, susceptibility to ambient interference, and high processing complexity. Contrarily, sensor-based human position identification techniques, such as those using depth cameras, audio signals, and inertial measurement units (IMUs), offer important benefits like high accuracy, low latency, and robustness to environmental factors [2]. These benefits support the quick advancement and extensive application of sensor-based activity recognition techniques. However, there also exist certain drawbacks to the sensor-centric approach. Firstly, sensor acquisition and integration amplify hardware costs and system intricacy, rendering it unsuitable for scenarios with constrained budgets or space requirements. Secondly, specific motions or actions might necessitate intricate algorithmic parsing, potentially leading to performance degradation. The processing and fusion of sensor data necessitate meticulous calibration and adjustments for varying sensor types, involving intricate technical challenges.

Human activity recognition entails monitoring and analyzing the body's movement status through the utilization of sensors, and automatically recognizing activity based on spatial or trajectory characteristics. The versatility of this approach allows for its application in various scenarios, bringing with it distinct advantages and specific conditions for implementation. Furthermore, the fusion of multiple sensor technologies enables a more thorough and comprehensive recognition and analysis of human activity, providing valuable insights into activity-related information. For example, placing sensors on different parts of the body (such as the wrist, waist, and ankle) can obtain more comprehensive information about human activity, helping to identify more complex behaviors such as bending, turning, etc. Human activity recognition can be achieved through different sensor-based methods. Inertial measurement units (IMUs) calculate body movements based on acceleration, angular velocity, and magnetic flux data from key body areas like wrists and knees [3]. Depth cameras, using infrared encoding, accurately recognize whole-body activities and joints [4]. Video sequence technology analyzes high-speed camera footage for activity recognition in sports and medical rehabilitation. Acoustic signal analysis estimates body activity by analyzing sound signals in the context of human–machine interactions and virtual reality scenarios. These methods enhance human–computer interactions, optimize sports training, aid medical rehabilitation, and advance virtual reality applications.

Although the IMU (inertial measurement unit) has been widely used in HAR, it still faces limitations in practical applications. The raw data from a single accelerometer is too simplistic to independently determine activity. Meanwhile, gyroscopes accumulate small errors over time with prolonged usage, leading the integrated values from the sensor to progressively deviate from true values. This phenomenon can result in the drift of integrated values over time, ultimately impacting the accuracy of applications [5]. As a result, many researchers have proposed multi-sensor fusion algorithms based on Kalman filters, combining attitude data from accelerometers, gyroscopes, and magnetometers. However, establishing stable observation equations for long-term monitoring is challenging due to the algorithm's inclusion of intricate mathematical principles, multi-dimensional state observation handling, and iterative updates. This complexity becomes particularly pronounced when dealing with high-dimensional state spaces or frequent data updates, resulting in escalated computational demands, and calculating measurement noise covariance and state noise covariance adds to the algorithm's complexity, leading to excessive overhead.

Therefore, this paper proposes a method of human pose angle extraction using an accelerometer and Euler angle calculation to recognize 19 kinds of human physical activities. Utilizing a sliding window technique, the method removes high-frequency noise to mitigate the impact of integration drift. Simultaneously, it introduces a module known as GAM to emphasize the temporal characteristics of sensor data, further counteracting integration drift. By adaptively adjusting model weights, it learns feature representations of the sensor data, reducing complex mathematical computations and enhancing recognition accuracy [6].

The contributions of the paper:

1. A human feature extraction model is introduced to convert multi-dimensional sensor data into one-dimensional features;
2. An approach for feature group division and classifier network construction is proposed to improve group correlation analysis and human action recognition accuracy;
3. The impact of the K-nearest neighbors algorithm (KNN) and sliding window size on evaluation results is examined;
4. The study investigates the influence of GAM, transformer block, and classifier block on the experimental accuracy.

The subsequent sections of this paper are organized as follows: The second section delineates recent advancements in the realm of human behavior recognition. Moving forward, the third section elucidates the transformation of Euler angles within human posture, alongside the methodology of integrating GAM-MLP information. Progressing further, the fourth section details the groundwork undertaken for the forthcoming experiments. In the subsequent fifth section, we present a comprehensive compilation of experimental findings, meticulously analyzed. Lastly, the sixth section encapsulates the entirety of the article through a concise summary, while also delving into an insightful analysis of potential avenues for future research opportunities.

2. Related Work

Recently, computer vision applied to image-based HAR has made significant progress. Bayraktar et al. [7]. have provided valuable insights into visual semantic analysis using deep neural networks and load analysis through sensors. This is noteworthy for harnessing computer vision and sensor load for human behavior recognition. However, visual-based HAR is susceptible to environmental factors such as lighting and angle variations. Therefore, the application of low-cost sensor-based HAR methods has become more prominent in this field. Yigit et al. [8]. proposed three comprehensive algorithms for low-cost variable stiffness robotics mechanisms. These algorithms enhance recognition accuracy and reduce costs in the adjustment of robot joint stiffness. In [9], an innovative approach was introduced to address the estimation of external forces and torques on elastic joints. This method leverages the inherent elastic characteristics of joints, eliminating the need for expensive sensors. These approaches provide some cost-effective solutions for joint sensor design in HAR.

When utilizing sensors for HAR, achieving a sufficient level of accuracy is crucial [10]. In scenarios where only a single sensor is employed for recognition, Anazco et al. developed a HAR system [11], in which individual IMUs are attached to the dominant wrist. They employed a variational autoencoder to automatically denoise the IMU signals, enhancing recognition accuracy. This approach achieved a recognition accuracy of 95.09% in everyday smartphone-based motion recognition. However, relying solely on sensors worn on the wrist has limitations when it comes to recognizing complex human gait behaviors. The choice of sensor placement significantly impacts the accuracy of human gait recognition. Abdelhafiz et al. devised a sensor selection approach using the feature selection criteria of maximum relevance and minimum redundancy. This method identifies optimal sensor placement for gait recognition [12]. Additionally, they implemented a two-layer classifier to differentiate interfered activities and incorporated physical features into the feature dictionary, thereby enhancing recognition accuracy.

Through collecting data using sensors placed at appropriate positions, neural network methods have demonstrated strong recognition capabilities when processing such data, Rivera et al. presented an approach for recognizing hand activities in daily life [13]. They proposed a deep autoencoder based on ARMA (Auto-Regressive Moving Average) and a deep recursive network using GRU (Gated Recurrent Unit). The ARMA-based deep autoencoder effectively denoises the unprocessed time-series signals, while the deep RNN-GRU utilizes the output from the encoder to identify seven hand gestures. This approach yielded a 12.8% increase in accuracy compared to traditional classifiers for

gesture recognition. Nevertheless, effective preprocessing techniques for sensor data remain essential to further enhance recognition accuracy.

Applying data preprocessing and model pretraining techniques often leads to remarkable results in dealing with complex human activity recognition tasks. Hashim et al. proposed a method to transform raw accelerometer and gyroscope sensor data into the visual domain [14]. To address the issue of high computational complexity during this process, they employed fine-tuning through pretraining a CNN and transfer learning. On several online human activity recognition datasets, they achieved a classification accuracy of 98.36%. Tahir [15] combined data preprocessing techniques with the primary domain features of human activities, such as time, frequency, wavelet, and time-frequency features. They employed a random forest classifier to monitor human body activities. Results on five commonly used HAR datasets demonstrated that their method exhibited a certain superiority compared to cutting-edge approaches. However, the challenge lies in how to process and fuse data from multiple sensors to achieve accurate human behavior recognition. This has emerged as a new challenge in the HAR field.

Fusing data from different sensors can enhance recognition accuracy. Chakraborty addressed the issue of recognition errors due to sensor placement and developed a heterogeneous sensor system [16]. This system utilized low-cost leg sensors and fingertip-based pulse sensors to acquire multimodal data. They employed a one-dimensional deep convolutional neural network for system performance evaluation. Through feature fusion, this approach achieved a 97% accuracy in recognizing the walking movements of the human body.

By monitoring and analyzing body movements through sensors and recognizing activities based on spatial and trajectory characteristics, the methods mentioned above open up versatile applications and offer valuable insights into activity-related information. For instance, the positioning of IMU sensors, data preprocessing techniques, and adjustments to network modules have all served as references in our approach presented in this paper. However, multi-sensor data fusion continues to face challenges associated with high complexity [17]. These challenges include overcoming the problem of the unified processing of multi-sensor data, concise attitude angle transformation, and behavioral feature extraction to avoid high complexity in the recognition process.

In order to address these issues, this paper introduces a human feature extraction model based on sensor data, aiming to collect and extract human posture angles to the maximum extent. The model uses dimensionality reduction technology to fuse the Euler angles of multiple parts of the human body into the overall pose angle, and utilizes the FFT algorithm for transforming time-domain characteristics into frequency-domain features, enhancing the accuracy of human activities recognition. Additionally, a novel technique called the Group Attention Module (GAM) is introduced, utilizing a multi-layer perceptron to share and fuse information among different features within the same group, effectively extracting the behavioral features of activities. This approach effectively improves the accuracy and robustness of HAR.

3. Methods

3.1. Enhancing the Human Pose Recognition Model Structure

The HAR system architecture involves data preprocessing to remove sensor noise and calculate Euler angles using roll, pitch, and yaw angles. Magnetometer and gyroscope data correct the initial Euler angles and reduce dimensionality, forming the attitude angle model. Feature extraction methods, including FFT, involve shifting to the frequency domain, reflecting human activity periodicity and frequency for practical significance [18]. Information fusion through a multi-layer perceptron (MLP) and a group attention module enhances feature fusion. Manual classification combined with conventional neural network techniques achieve accurate HAR. The fully connected and SoftMax layers produce the final classification results, enabling precise human activity recognition. Figure 1 illustrates

the identification process of the hybrid human pose recognition method using an MLP neural network and Euler angle extraction based on IMU sensors.

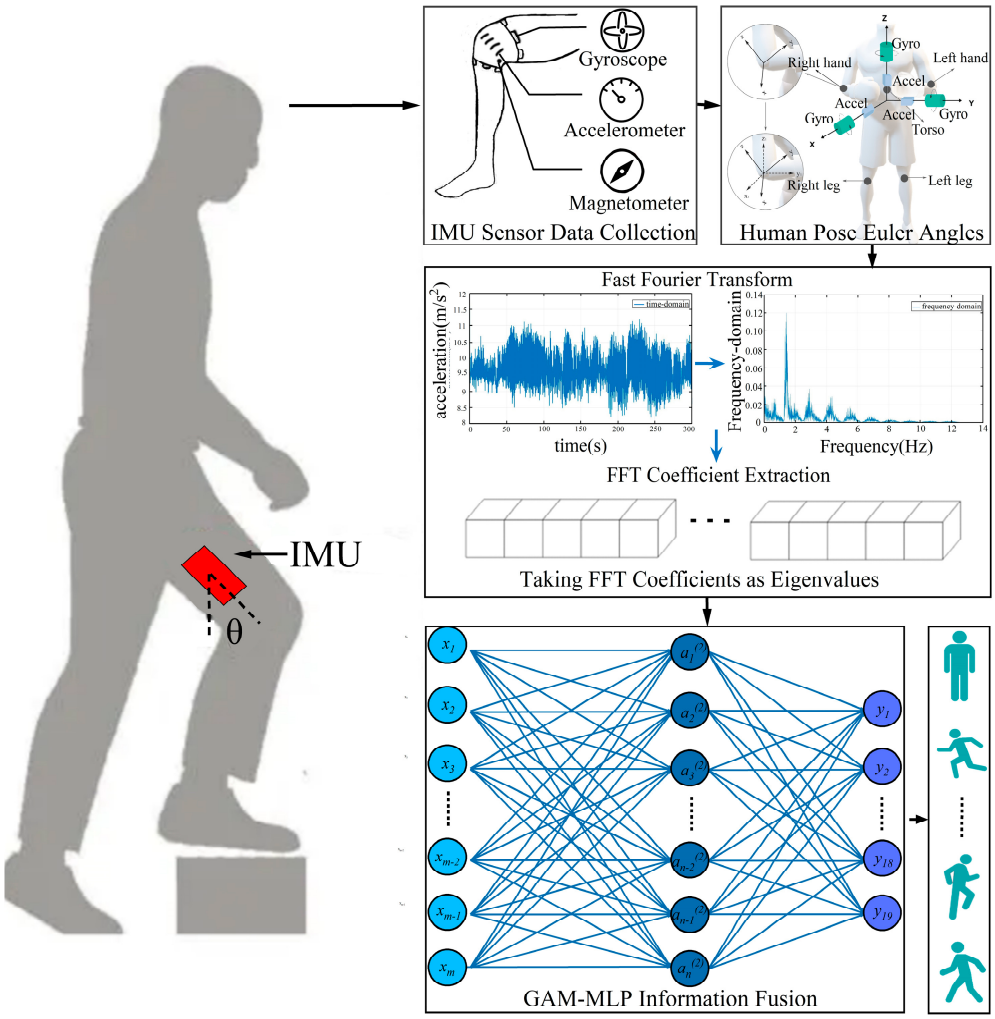


Figure 1. Identification process of hybrid human activity recognition method using MLP neural network and Euler angle extraction based on IMU sensors.

3.2. Extraction of Human Pose Euler Angles

3.2.1. Initial Attitude Angle Calculation

We use Euler angles to describe the orientation of several body parts [19–21], including the torso, left hand, right hand, left leg, and right leg, in order to calculate the initial torso rolling angle, pitch angle, and yaw angle. The acceleration vectors of various body parts must be converted into a single coordinate system in order to calculate the Euler angles with accuracy, assuring consistency in angle calculation and dimensionality reduction.

To get the torso's Euler angles, the following procedures are conducted, assuming the common coordinate system is represented by xyz:

First of all, the roll angle of the torso, indicated as $\varnothing_t^{torso}$, is calculated as follows:

$$\varnothing_t^{torso} = tan^{-1}\frac{a_{y,t}^{torso}}{a_{z,t}^{torso}} \quad (1)$$

where, in the common coordinate system, $a_{y,t}^{torso}$ denotes the acceleration along the Y-axis of the acceleration vector and $a_{z,t}^{torso}$ signifies the acceleration along the Z-axis. Furthermore, the pitch angle of the torso θ_t^{torso} is computed using the formula:

$$\theta_t^{torso} = tan^{-1} \frac{-a_{x,t}^{torso}}{\sqrt{\left(a_{y,t}^{torso}\right)^2 + \left(a_{z,t}^{torso}\right)^2}} \tag{2}$$

In this equation, the term, $a_{x,t}^{torso}$ denotes the acceleration of the torso's acceleration vector in the common coordinate system along the X-axis. The yaw angle of the torso, shown by the symbol φ_t^{torso}, is then calculated. The initial yaw angle is set to zero during the computation since accelerometers can only detect an object's acceleration in relation to a fixed coordinate system and not its absolute direction:

$$\varphi_t^{torso} = 0 \tag{3}$$

Once the Euler angles of the torso in the common coordinate system have been computed, the Euler angles for the other body parts can be calculated using specific methods tailored to each part. For instance, the acceleration vector of the right hand must be rotated so that it is perpendicular to the xz plane of the torso in order to calculate the Euler angles of the right hand. In order to accurately calculate the Euler angles in the torso coordinate system, this rotation seeks to remove the right hand's projection in the xz plane of the torso. Figure 2 is right hand coordinate transformation, gyroscope, and accelerometer schematic for the human body model, in which we describe the 3D human model.

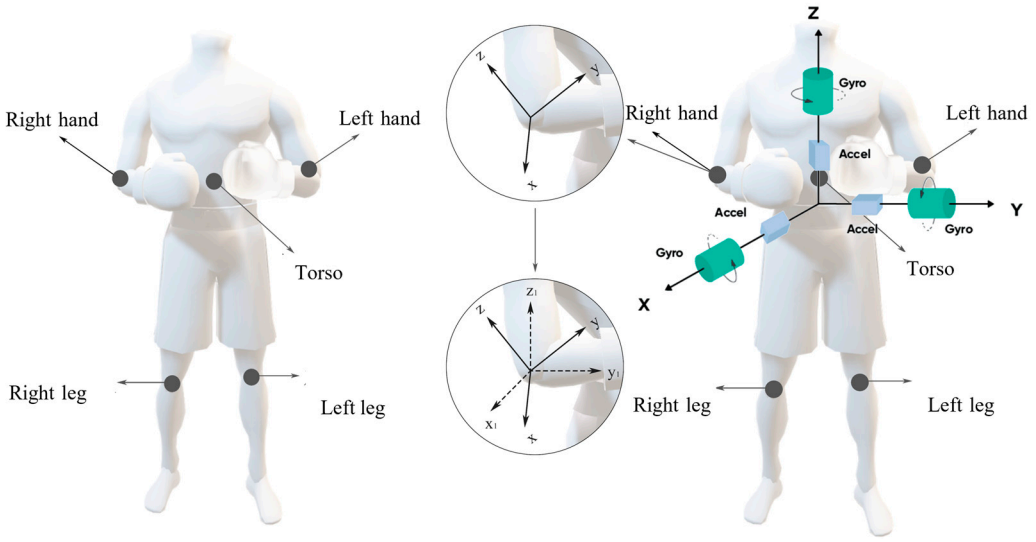


Figure 2. Right hand coordinate transformation, gyroscope, and accelerometer schematic for the human body model.

The process involves first calculating the rotation angle of the right hand vector, denoted as $\alpha_t^{rightarm}$, relative to the coordinate system of the torso:

$$\alpha_t^{rightarm} = tan^{-1} \frac{-a_{z,t}^{torso}}{a_{y,t}^{torso}} \tag{4}$$

where, $a_{z,t}^{torso}$ represents the acceleration along the Z-axis of the acceleration vector of the right hand position in the common coordinate system, and $a_{y,t}^{torso}$ denotes the acceleration along the Y-axis.

Next, the acceleration vector of the right hand after rotation, represented as $\alpha_t^{rightarm}$, is obtained by rotating the original acceleration vector $\vec{\alpha}_t^{rightarm}$ about the x-axis of the torso coordinate system, using the rotation matrix $R_x(\alpha_t^{rightarm})$ [22,23]:

$$\vec{\alpha}_t^{rightarm} = R_x(\alpha_t^{rightarm})\vec{\alpha}_t^{rightarm} \tag{5}$$

$$R_x(\alpha_t^{rightarm}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha_t^{rightarm}) & -sin(\alpha_t^{rightarm}) \\ 0 & sin(\alpha_t^{rightarm}) & cos(\alpha_t^{rightarm}) \end{bmatrix} \tag{6}$$

To begin with, the magnetometer data is utilized to compensate for the yaw angle in the original attitude angle by determining the direction of the horizontal plane of Earth's magnetic field. Since the magnetic field component perpendicular to the horizontal plane does not impact the calculation of the yaw angle, the orientation of the magnetic field within the horizontal plane is employed for compensation. The formula for compensating the yaw angle is given as follows:

$$\varphi_t = 0 + atan2(m_y, m_x) \tag{7}$$

Additionally, the original Euler angles are compensated using gyroscope data from various body parts, following a similar calculation approach. The compensation method for the torso's original Euler angle using gyroscope data is described below.

Gyroscope data can be integrated to compensate for the torso's Euler angles. Specifically, the gyroscope data compensation formula is as follows:

$$\begin{cases} roll_{comp} = roll_{prev} + \frac{gyro_x \cdot dt}{cos(pitch_{prev}) \cdot cos(yaw_{prev})} \\ pitch_{comp} = pitch_{prew} + \frac{gyro_y \cdot dt}{cos(roll_{prev}) \cdot cos(yaw_{prev})} \\ yaw_{comp} = yaw_{prew} + \frac{gyro_z \cdot dt}{cos(roll_{prev}) \cdot cos(pitch_{prev})} \end{cases} \tag{8}$$

where, dt represents the sampling time interval, $roll_{prev}$, $pitch_{prew}$, and yaw_{prew} are the Euler angles of the torso at the previous sampling moment. These formulas employ the integration of gyroscope data to update the current Euler angles of the torso, thereby compensating for the gyroscope data.

Generally, the Euler angle principle is an early method for attitude determination due to its straightforward physical interpretation. However, it involves complex operations with many trigonometric calculations. To simplify the practical application process, we introduce the quaternion method as an optimization of the Euler angle approach.

3.2.2. Euler Angle Correction for Quaternion and Rodriguez Parameters

Given the roll $\varnothing$, pitch θ, and yaw φ Euler angles for each body part, these can be converted into rotation quaternions. The quaternion representing the orientation of a body part in the common coordinate system would be in the form:

$$q = cos\left(\frac{\varnothing}{2}\right)cos\left(\frac{\theta}{2}\right)cos\left(\frac{\varphi}{2}\right) + sin\left(\frac{\varnothing}{2}\right)sin\left(\frac{\theta}{2}\right)sin\left(\frac{\varphi}{2}\right)i + sin\left(\frac{\varnothing}{2}\right)cos\left(\frac{\theta}{2}\right)cos\left(\frac{\varphi}{2}\right)j + cos\left(\frac{\varnothing}{2}\right)sin\left(\frac{\theta}{2}\right)cos\left(\frac{\varphi}{2}\right)k \tag{9}$$

$$\begin{cases} \varnothing = atan2(2(q_0q_3 + q_1q_2), 1 - (q_2{}^2 + q_3{}^2)) \\ \theta = asin(2(q_0q_2 + q_1q_3)) \\ \varphi = atan2(2(q_0q_1 + q_2q_3), 1 - (q_1{}^2 + q_2{}^2)) \end{cases} \tag{10}$$

where, q_0, q_1, q_2, and q_3 are the components of the quaternion.

Compared to the Euler angle method, the quaternion description method further simplifies the calculation process and effectively avoids the singularity problem of the Euler angle due to a trigonometric function operation, but quaternions possess redundant parameters and quaternion normalization constraints. Quaternion tracing was compared to the Euler angle method singularity problem of the angle function operation. Lastly, to facilitate the extraction of human activity posture features, a dimensionality reduction technique is applied to the optimized Euler angles obtained through the above steps. The resulting reduced Euler angles are defined as the attitude angle, denoted as attitude. The formula for computing the attitude angle is as follows:

$$attitude = arccos(cos\varnothing cos\theta cos\varphi + sin\varnothing sin\theta sin\varphi) \tag{11}$$

where, $\varnothing$, θ, and φ are the roll, pitch, and yaw angles, respectively. This formulation combines the three-dimensional Euler angles into a single one-dimensional attitude angle, providing a concise representation of the overall posture.

3.2.3. Attitude Angle Calculation Algorithm Design

A new method of human attitude angle calculation based on multiple attitude parameters compensating each other is proposed, which makes up for the redundant parameters and the singular values of trigonometric functions. The attitude angle calculation procedure begins with the initialization process and unit conversion of acceleration data (lines 1–2). Subsequently, Euler angles are computed based on the sampled frequency-processing sequence data, with compensation provided through gyroscope and accelerometer data (lines 3–12). Following this, the Euler angle vector is transformed into a rotation matrix, and from there into a quaternion and Rodrigues parameters (lines 13–15). Finally, the ultimate pose angles are calculated and displayed (lines 16–19).

Algorithm 1: The human attitude angle calculation method based on multiple attitude parameters

Input: Sample time series T, Acceleration data A_{data} (a_x, a_y, a_z), Gyroscope data G_{data}, Magnetometer data M_{data};
Output: Attitude Angle *attitude*;
(1) Initialization A_{data}, M_{data}, Quaternion Q and Rodrigues Parameters r;
(2) Conversion unit of accelerometer data to the acceleration of gravity
(3) **for** i = 0,0.04,0.08…T **do**:
(4) Calculate roll angle $\varnothing_t$ and pitch angle θ_t
(5) Compensate for M_{data} to correct Attitude Angle with magnetometer
(6) Calculate yaw angle φ_t
(7) Convert angles to radians
(8) Update Euler Angle vector E_{angle}
(9) Calculate the Gyroscope Euler Angle variation:
(10) $G_{change} = G_{data} * d_t$
(11) Compensate the Euler Angle C_{comp}:
(12) $C_{comp} = E_{angle} + G_{data}$;
(13) Transform Euler Angle vector to rotation matrix
(14) Convert Euler Angle vector to quaternion
(15) Convert quaternion to Rodrigues Parameters
(16) Calculate Attitude Angle *attitude*
(17) Convert radians to angles
(18) **end for**
(19) Output the final Attitude Angle *attitude*

3.3. Human Pose Feature Extraction

FFT is applied to the data [24], yielding the frequency-domain conversion results depicted in Figure 3.

$$X[k] = \sum_{n=0}^{N-1} x[n] \times e^{\frac{-j2\pi kn}{N}} \tag{12}$$

where, $X[k]$ represents the frequency-domain signal, $x[n]$ denotes the denoised time-domain signal of length N, N represents the number of sampled data points, and j represents the imaginary unit $(\sqrt{-1})$.

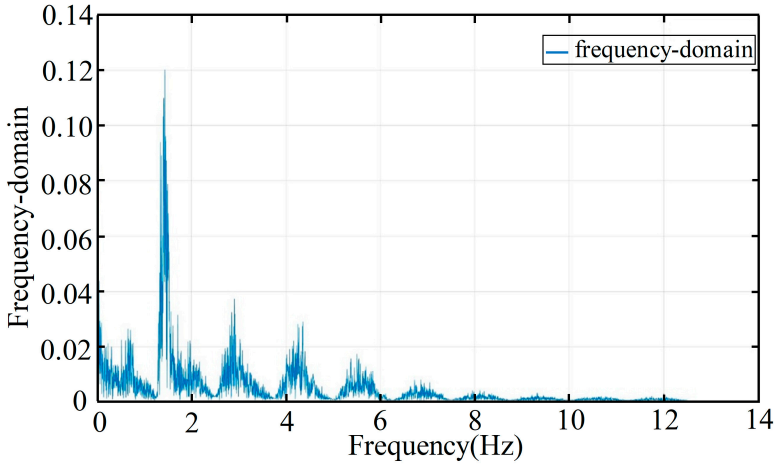


Figure 3. The processing graph for FFT feature extraction.

The first 45 FFT coefficients are then selectively extracted as the 45 eigenvalues, which represent the human activity.

$$X_{eigenvalues} = [X[0], X[1], \cdots, X[44]] \tag{13}$$

These eigenvalues are subsequently harnessed for deep learning-based human body activity recognition.

3.4. GAM-MLP Information Fusion

3.4.1. Human Pose Feature Information Fusion

In order to better process the spatiotemporal features of IMU sensor data, the neural network model that is suggested in this paper incorporates information fusion [25], attention evaluation [26], and classification recognition [27]. Figure 4 depicts the model architecture of GAM-MLP and the process of action classification. The data of different sensors in each part of the human body are put into the GAM-MLP model as one-dimensional features to recognize human activities, and finally the recognition results are obtained. The data format represented by (, 64) in the figure represents (batch_size, data dimension), due to batch_ size is not fixed and is therefore represented as blank.

GAM-MLP's primary objective is to extract posture features of the utmost significance for action recognition from a 20-dimensional dataset. This extraction process is intended to facilitate effective information fusion and utilize these features as the primary foundation for classification. To accomplish this objective, the paper introduces a component called the GAM.

In conventional neural network classification methodologies, each pose feature is uniformly trained, with their significance determined solely by the weights assigned to individual neurons. However, this method suffers from inadequate communication among distinct pose features, ultimately yielding less impactful results. This paper introduces a novel approach that combines manual partitioning and automatic learning. This hybrid approach empowers the neural network model to allocate varying degrees of importance to different categories of pose features, ultimately resulting in an enhanced classification accuracy.

In the GAM, the pose features are manually categorized into distinct groups. During the manual grouping process, following feature extraction via FFT, it becomes evident that there exist resemblances in the characteristics among features associated with various

activities. For instance, the frequency features of "Sitting" and "Standing" are characterized by frequencies of less than 4 Hz and exhibit a unimodal pattern. On the other hand, activities such as "Exercising on a stepper" and "Playing Basketball" demonstrate characteristic frequencies reaching approximately 10 Hz, marked by multiple peaks and substantial fluctuations. Consequently, human motion activities can be effectively classified into dynamic and static groups. The static group conforms to low-frequency attributes with a single wave peak, while the dynamic group encompasses high-frequency elements, multiple peaks, and significant fluctuations. As depicted in the accompanying Figure 5, we provide a visual comparison of feature images extracted from several activity features.

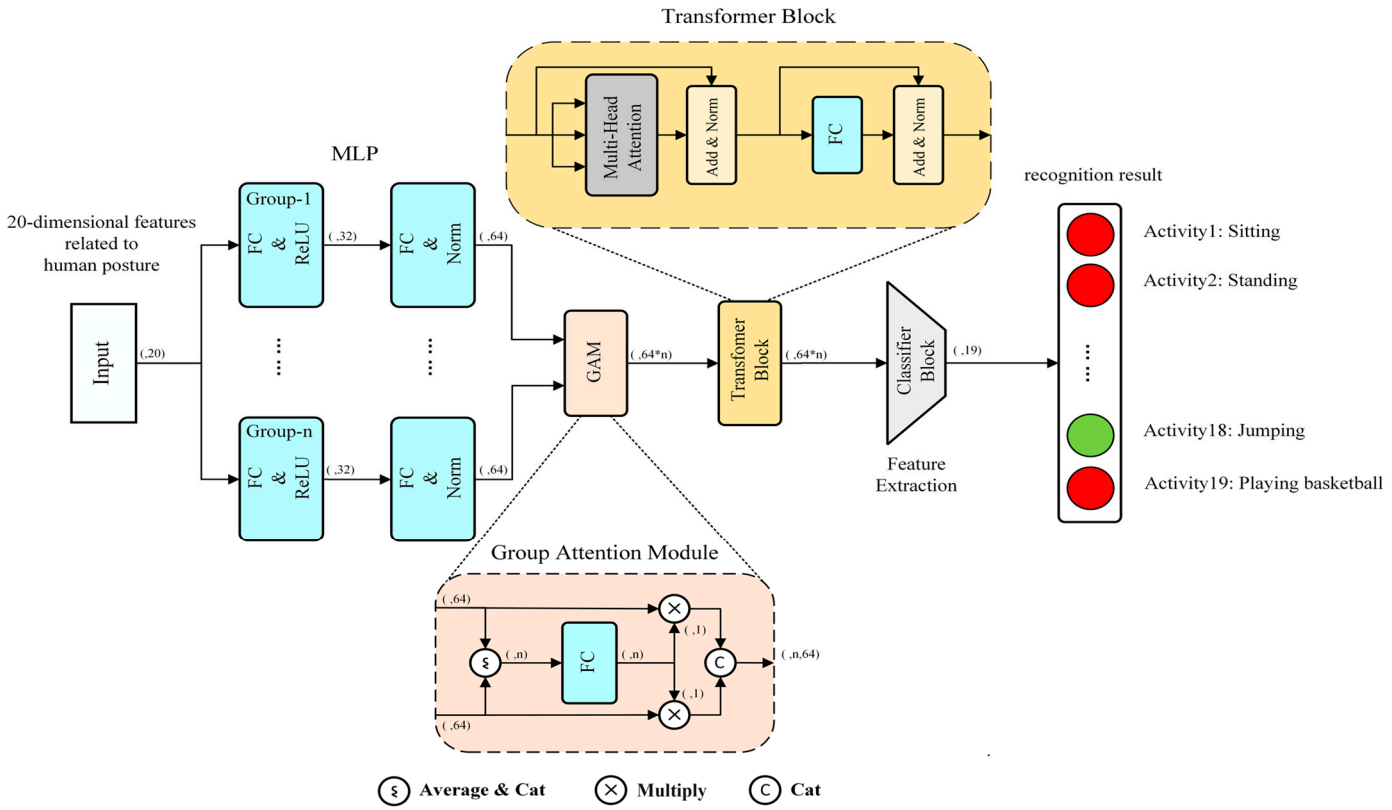
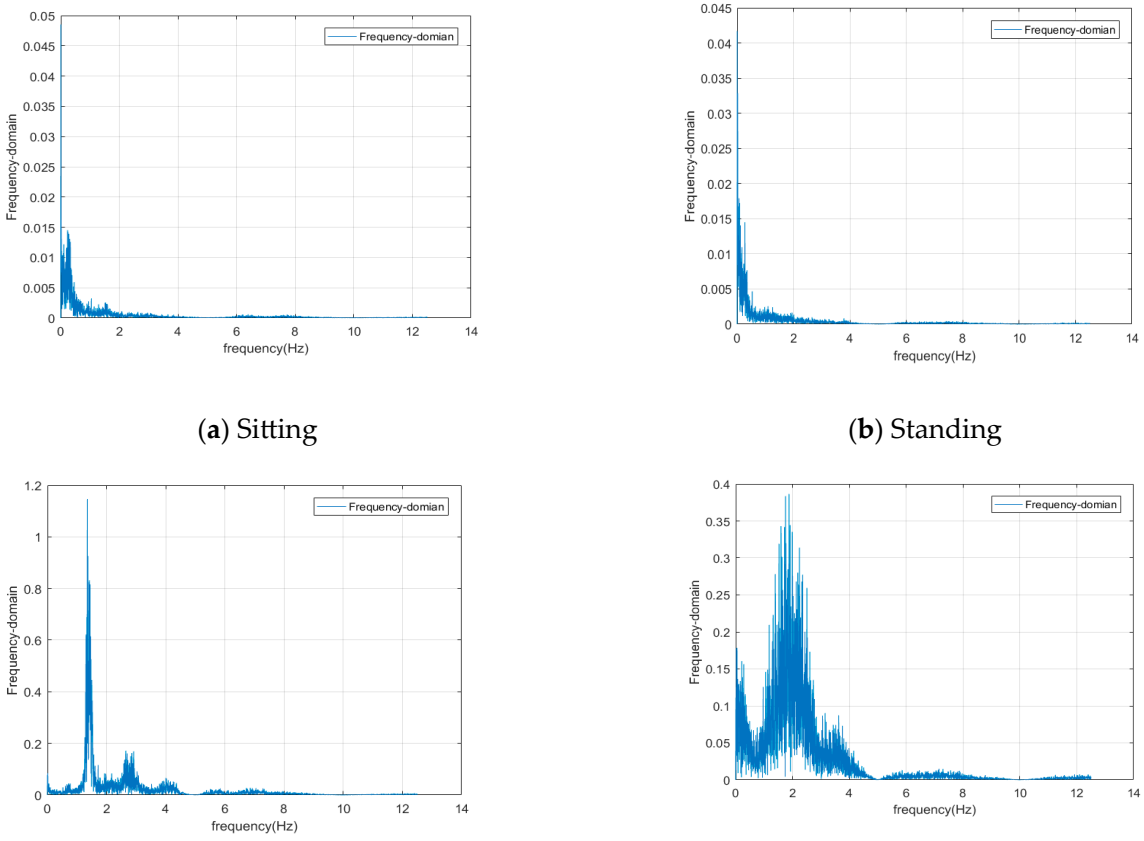


Figure 4. The model architecture of GAM-MLP and the process of action classification.

Furthermore, adhering to the principles of the periodicity inherent in dynamic motion, the dynamic group can be further categorized into two subgroups: periodic dynamic activity and random dynamic activity. Within the periodic dynamic activity category, distinctions can be made based on the unique patterns and trends of peak characteristics. To summarize, the criteria and guidelines for manual partitioning can be systematically delineated, progressing from higher-order groupings to lower-order ones, in accordance with the specific characteristics of the dataset. The schematic representation of the group attention module is illustrated in Figure 6. The blocks with different colors in the figure represent the weights calculated by different groups.

To address the inherent nonlinear relationships and patterns within both the dynamic and static groups of features, this paper introduces the "Transform Block", which combines elements from MLP and the Transformer model. Both MLP and the Transform Block are integral components of deep learning models, commonly employed in tandem to handle intricate datasets and tasks.



(**a**) Sitting (**b**) Standing

(**c**) Exercising on a stepper (**d**) Playing basketball

Figure 5. Different activity features of FFT feature extraction.

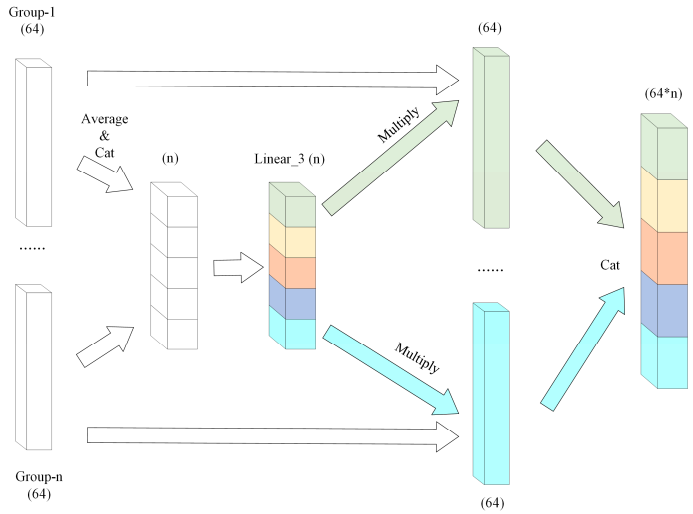


Figure 6. Flow chart of action classification and recognition with group attention module.

MLP serves the purpose of introducing nonlinearity into the model, while the Transform Block, a crucial part of the Transformer architecture, is utilized for effectively modeling sequential data. This integration is geared towards enhancing the model's capacity for representation learning and nonlinear modeling, thereby enabling it to adaptively generate distinct weights between modules and effectively capture the intricate relationships and patterns present in the data.

The Multi-Layer Perceptron (MLP) serves as a conduit for interchanging and amalgamating distinct features within the same group. It encompasses a series of layers including

the fully connected, activation, and standardization layers [28–30]. The training process employs backpropagation, iteratively adjusting weights and biases. This fosters incremental learning and the integration of pertinent information across features in the same group.

The fully connected layer embodies input and output neurons. Given the ith input, x_i, and the jth output, y_j, their relationship is governed by the weight, w_{ij}, and bias, b_j, represented as

$$y_j = \sum_{i=1}^{10} w_{ij} x_i + b_j \tag{14}$$

Two fully connected layers are followed by the activation layer and the standardization layer, respectively, to achieve the nonlinearity of network representation and improve training speed, generalization ability, and robustness, while reducing overfitting, gradient disappearance or gradient explosion problems. For the activation layer, the Rectified Linear Unit (ReLU) is selected, which is more efficient and safer than other activation functions while making the input of neurons a nonlinear transformation. Its function expression is shown as follows:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{15}$$

where max represents taking the larger value of the two numbers, and 0 is the cutoff point of ReLU, that is, when $x < 0$, the value of ReLU is 0.

For the normalization layer, we utilize batch normalization. The characteristics of each sample, $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_m^{(i)}\right)$, where m indicates feature count, are transformed according to

$$\text{Mean} : \mu_B = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{16}$$

$$\text{Variance} : \sigma_B^2 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 \tag{17}$$

$$\text{Normalization} : \hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{18}$$

$$\text{Output} : y_i = \gamma \hat{x}_i + \beta \tag{19}$$

where μ_B and σ_B^2 denote batch mean and variance, respectively, $\hat{x}_i$ is the normalized feature, γ and β are learnable parameters for scaling and shifting, and ϵ prevents division by zero. This process standardizes data, maintains distribution information, and optimizes data representation.

The synergetic integration of these layers fortifies the MLP's capacity to comprehend and amalgamate interrelationships among features within the same group. This amalgamation enhances activity recognition accuracy through a robust and sophisticated learning process.

Subsequently, we employ transformer blocks to enhance attention evaluation and sequential modeling, leveraging the self-attention mechanism to improve the model's performance. Specifically, the transformer block facilitates establishing global dependencies across various points within the input sequence, enabling the capture of longer-range semantic dependencies and enhancing the model's efficacy in human motion recognition tasks.

Transformer blocks typically comprise several sub-layers, including a self-attention layer, a fully connected layer, and a normalization layer. The self-attention layer calculates attention weights at different positions within the input sequence, facilitating the correlation of distinct parts of the input sequence. The fully connected layer further processes and transforms the outcome from the self-attention layer. After the self-attention and fully connected layers, the input and output are combined through a residual connection, while the neural network training process is expedited via batch normalization. This helps to prevent issues like gradient vanishing or explosion, ultimately improving the model's generalization capabilities. Multiple transformer blocks can be stacked together to construct a deeper neural network, further enhancing the model's overall performance.

Self-attention is a core component of the transformer model, playing a crucial role in machine learning tasks in natural language processing and other domains. It can associate and interact with information from different positions in the input sequence to capture the long-term dependencies in the sequence.

The self-attention [31] calculation processes a sequence $(x_1, x_2, \ldots, x_n)$, with each x_i being a vector. The self-attention layer yields an output sequence $(y_1, y_2, \ldots, y_n)$, where each y_i, is also a vector. The self-attention mechanism can be explained as follows:

$$\text{SelfAttention}(X) = \text{softmax}\left(\frac{X \cdot W_Q (X \cdot W_K)^T}{\sqrt{d_k}}\right) X \cdot W_V \tag{20}$$

where $X \in R^{n \times d}$ represents the input sequence, W_Q, W_K, and $W_V \in R^{d \times d_k}$ are weight matrices that map input vectors to query, key, and value vectors. The output of self-attention is a weighted average of input position vectors, with weights calculated through similarity (dot product) between position vectors after scaling and Softmax normalization.

Residual connection addresses gradient vanishing in deep neural network training by allowing input data to flow directly to subsequent layers during forward propagation. This preserves input information, alleviating information loss and distortion. This concept applies beyond convolutional neural networks and enhances various neural network structures' training.

Finally, the weighted feature group representations were added together to obtain the final attention fusion representation. The proposed group attention module and transformer module empower the neural network to focus on different input parts, enhancing adaptability across tasks and scenes. In activity recognition, it aids the network in emphasizing activity-related input information, improving accuracy and generalization.

3.4.2. Activity Classification and Recognition

The GAM-MLP model concludes with a fully connected layer followed by a SoftMax layer for classification purposes [32]. The diagram illustrating the classification and recognition process in the MLP of layer plus SoftMax layer is shown in Figure 7 below.

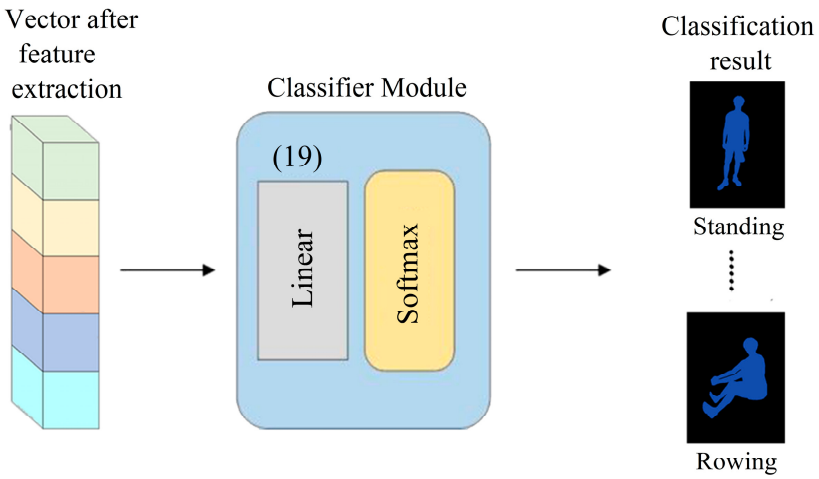


Figure 7. Flow chart depicting the classification and recognition in the MLP of layer plus SoftMax layer.

The function of the SoftMax layer is to transform neural network output scores into a probability space, ensuring category probabilities range from 0 to 1, summing to 1. For neural network output scores $z = (z_1, z_2, \ldots, z_n)$, the Softmax layer's output is given by

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{21}$$

Here, e is the base of the natural logarithm, and y_i represents the probability for the ith class.

4. HAR Datasets and Experiment Settings

4.1. Experimental Environment and Data Acquisition

For human body pose recognition using IMU sensors, we have meticulously assembled a specialized hardware and software configuration. The system features an Intel Core i9-12900K CPU at 3.50 GHz and an Nvidia GeForce RTX 3080Ti GPU with 12 GB VRAM, ensuring high-speed computation for accurate pose analysis. With 64 GB DDR4 3000 RAM, memory-intensive operations are efficiently handled. TensorFlow 2.5 serves as the deep learning framework, supported by Python 3.9 as the primary development language. PyCharm 2020.1 IDE offers a user-friendly platform for coding and debugging. Windows 10 Professional OS provides stability and reliability for precise pose recognition. This setup optimizes IMU sensor capabilities, benefiting sports biomechanics, healthcare, and animation applications. The relevant configurations of the experiment are shown in Table 1.

Table 1. High-performance AI workstation for advanced deep learning applications.

Name	Specific Configuration
CPU	Intel Core i9-12900K@3.50 GHZ
Graphics card	Nvidia Geforce RTX 3080Ti (12 GB) GPU
Memory	64 GB DDR4 3000
Deep learning framework	TensorFlow 2.5
Development language	Python 3.9
Developing an IDE	Pycharm 2020.1
Operating system	Window10 Professional

Sensor-based human activity recognition is a sophisticated technology that utilizes sensor technology to monitor and analyze the spatial positioning and motion characteristics of the human body. It automatically recognizes and interprets various human activities, making it applicable in diverse scenarios. We employ an inertial measurement unit (IMU) sensor technology to achieve a more comprehensive and in-depth recognition and analysis of human activity information.

The datasets used in the experiment are the PAMAP2 dataset and the MultiportGAM dataset, where the PAMAP2 dataset is a public dataset that can be publicly obtained. The MultiportGAM dataset is our self-made dataset, which includes 19 different types of activities and generates a total of 11,034 sample data.

We invited 10 volunteers to collect sensor data for the MultiportGAM dataset. To ensure the accurate collection of human activity data, we utilized specific sensors for different body parts. An accelerometer is employed to measure the acceleration of the human torso, left hand, right hand, left leg, and right leg. Additionally, a magnetometer is used to measure the magnetic field strength and direction of these body parts. Furthermore, a gyroscope is utilized to measure the angular velocity of the aforementioned body segments. The eight types of daily activities collected are shown in Figure 8. By compensating for the data from the magnetometer and gyroscope, more precise localization of human joints is achieved, enhancing the accuracy of human activity identification.

The sampling frequency of the sensor is 30 Hz, and the number of collected action types is 19. Each volunteer completed these actions within 1–2 h, with an average duration of 3–5 min for each action. In order to keep the behavioral data collected by sensors closer to the real-world human behavior, some specific actions such as cycling, shooting, paddling, etc., in Figure 8 are collected using sensors with the help of relevant equipment. Daily actions such as standing, sitting, walking, going upstairs, lying down, etc., are carried out in an orderly manner under indoor conditions to ensure the stability of the collection environment and avoid magnetic field interference. In addition, each row of data includes sensors for calibration to eliminate bias and improve measurement

accuracy. The collected raw action data includes action labels and timestamps for subsequent data analysis and alignment.

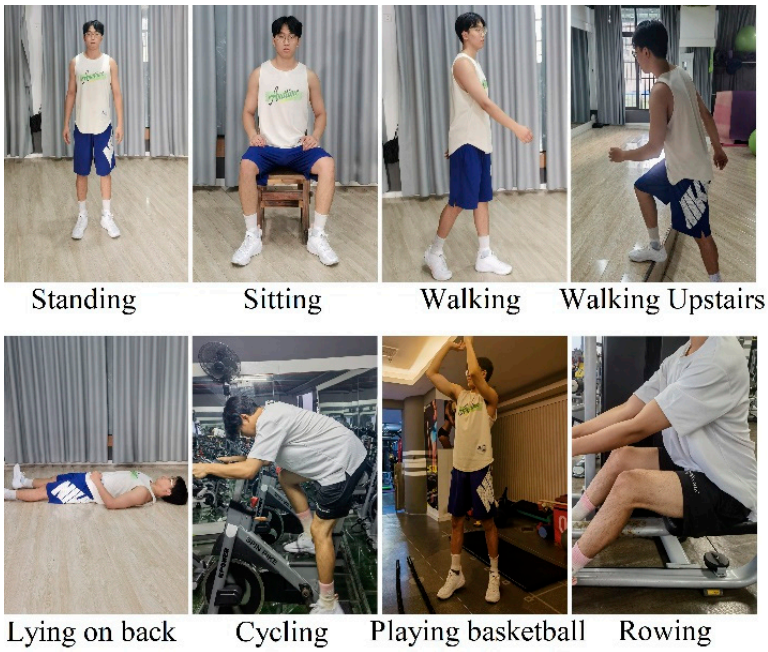


Figure 8. MultiportGAM dataset for sensors-based tester.

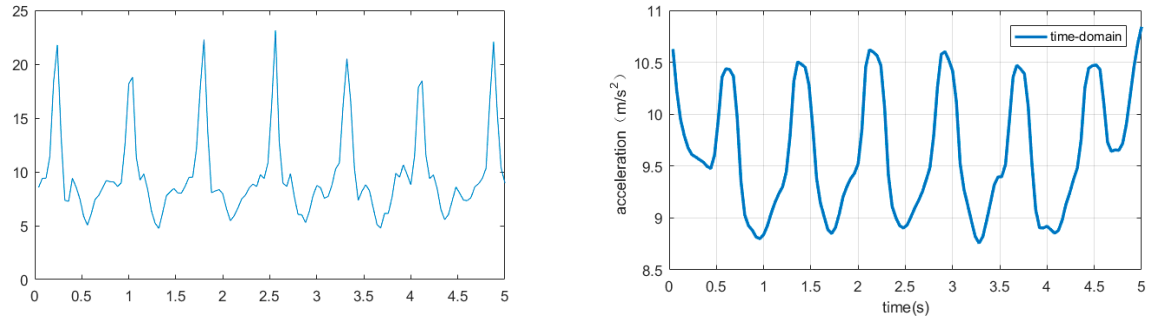
Table 2 records the basic information of the two datasets, including activity category, sample size, and sensor type. Compared to the PAMAP2 dataset, the MultiportGAM dataset has fewer samples, but it encompasses a greater variety of sensor placement. Both datasets capture common human activities from daily life.

Table 2. Comprehensive human activity recognition datasets with sensor information.

Dataset	Action Categories	Sample Size	Sensor Type	Acquisition Location
PAMAP2 Dataset	18	65,052	inertial measurement units, acceleration sensors, magnetometer, gyroscope	Ankle Chest Wrist
MultiportGAM Dataset	19	11,034	accelerometer, magnetometer, spirometer	Torso Left Arm Right Arm Left Leg Right Leg

4.2. Sliding Window Segmentation Signal Processing

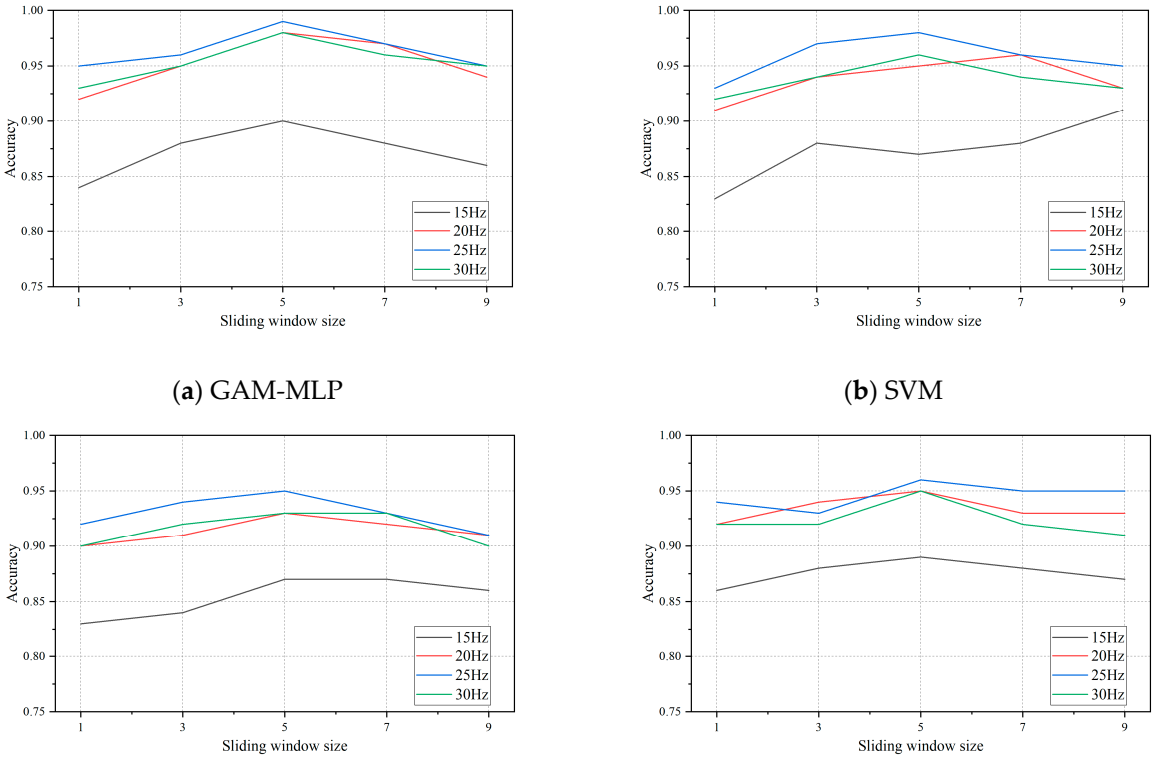
Continuous data signals gathered by sensors in practical applications are frequently plagued by noise and abrupt interference, which can dramatically reduce measurement accuracy. In the process of human activity recognition, the data of the sensor will present certain gait characteristics when the human body performs repetitive unit actions. After noise reduction via sliding window, the smoothness of the data time series is preserved and the motion characteristics are highlighted, as shown in Figure 9. The implementation of a sliding window denoising approach improves the accuracy and efficiency of the process.



(**a**) (**b**)

Figure 9. Exercising within a sample 5 s on a stepper acceleration time-domain diagram; (**a**) description of the original data before sliding window denoising processing; (**b**) description of the data after sliding window denoising processing.

To identify the optimal sliding window size and sampling frequency values, noise reduction is applied to the specific torso data adopted by the tester. The reason for adopting the torso data here is that when the human body performs periodic activities, the torso differs from the rest of the body in typical general gait characteristics. At the same time, in this study, the sliding window size was varied, ranging from 1 to 9 s. The sampling frequency was adjusted within the range of 15 to 30 Hz, with a step size of 5 Hz. The average accuracy was used as the scoring Index to solve the optimal value of the sliding window size and sampling frequency, and the evaluation scores of different machine learning algorithms were obtained as shown in Figure 10.



(**a**) GAM-MLP (**b**) SVM

(**c**) IN (**d**) LSTM-CNN

Figure 10. Accuracy scores obtained with different sliding window sizes, sampling frequencies, and machine learning algorithms.

From the results, it can be found that the different algorithms have a peak at a window size of 5 s and a frequency of 25 Hz which indicates that choosing one particular algorithm

over another does not make a difference. Also, an accuracy of 0.92–0.99 is good enough for our purposes. Therefore, these values will be used for further experiments and discussion.

5. Result and Analysis

5.1. Accuracy and Loss of the 10-Fold Cross Validation on Both Training and Test Sets

The Euler angle data collected by sensors and calculated using Algorithm 1 is used as input for the information fusion experiment after separating the training and test sets. Figure 11 shows that the upgraded GAM-MLP method's Train acc and Test acc curves exhibit a better match.

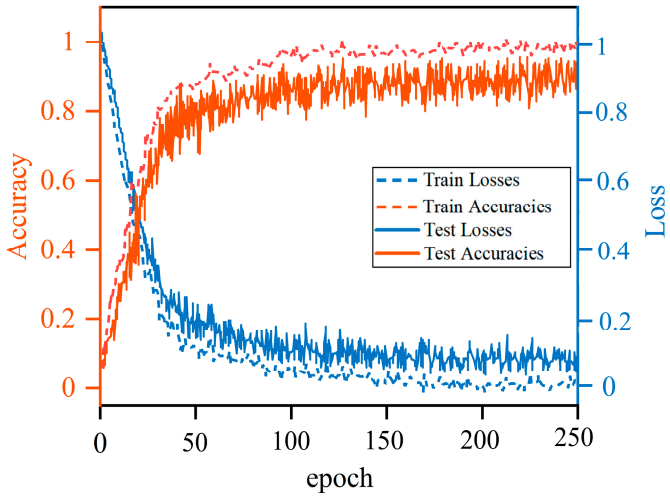


Figure 11. Accuracy and Loss of GAM-MLP on both training and test sets.

Additionally, the accuracy of the test set and training set stabilize as the number of iteration rounds approaches 120, demonstrating outstanding robustness and generalization capabilities.

From the accuracy curves of the training and testing sets in Figure 11, it can be seen that with the increase of iteration rounds, the curves representing the accuracy of the training and testing sets steadily increase and maintain alignment. This indicates that GAM-MLP performs as well as the training data when facing new data, demonstrating good generalization ability.

5.2. Performance on Different Datasets

We proceed to compare our hybrid human activity recognition method with both baseline techniques and several cutting-edge techniques on the MultiportGAM and PAMAP2 dataset. We evaluate each model's performance using three key metrics: accuracy, recall, and F1-score. Accuracy can intuitively measure the model's ability to accurately predict human activity categories. The recall rate represents the proportion of the number of behavior categories correctly predicted by the model to the actual number of human behaviors. A high recall rate means that the model can more comprehensively detect human behavior and reduce missed recognition. The F1-score provides a more comprehensive evaluation of the model's performance in human behavior recognition tasks. A model with a high F1-score demonstrates strong performance in terms of both precision and recall, being able to more accurately recognize human behavior and possessing good overall performance. This comprehensive analysis aims to uncover the effectiveness of our hybrid human pose recognition approach in the realm of human activity recognition.

The results of our comparisons on the MultiportGAM dataset are presented in Table 3, offering a clear perspective on the respective performances. Similarly, the comparison outcomes with the PAMAP2 dataset are summarized in Table 4. By examining the met-

rics across these datasets, we gain valuable insights into how our hybrid human pose recognition method performs in comparison to other established and advanced techniques.

Table 3. Comparison of different methods for the MultiportGAM dataset.

Method	Accuracy (%)	Recall (%)	F1-Score (%)
GAM-MLP	96.13	96.12	96.13
SVM	82.70	82.70	82.69
DCL [33]	92.11	92.10	92.11
IN [34]	92.72	92.72	92.71
LSTM-CNN [35]	94.23	94.10	94.17

Table 4. Comparison of different methods for the PAMAP2 dataset.

Method	Accuracy (%)	Recall (%)	F1-Score (%)
GAM-MLP	93.96	93.89	93.91
SVM	82.84	82.43	82.58
CNN-M [36]	93.74	93.28	93.85
LSTM-CNN	92.63	92.61	92.89
FE-CNN [37]	91.66	91.43	91.40
DCL	92.49	92.42	92.30
CE-HAR [38]	92.14	92.43	92.18
IN	91.77	91.76	91.47
TL-HAR [39]	92.33	91.83	92.08
ConvAE-LSTM [40]	94.33	-	94.46

Based on the comparative results, it is evident that leveraging the human pose angle model constructed from raw sensor data and the GAM-MLP model's attention mechanism, we achieved remarkable performance on the MultiportGAM Dataset. Our hybrid approach attained an accuracy of 96.13%, a recall rate of 96.12%, and an F1-score of 96.13%. In contrast to the baseline methods, our approach demonstrated notable improvements of 13.43%, 13.42%, and 13.44% in the respective accuracy, recall, and F1-score metrics.

The hybrid approach surpasses the currently established cutting-edge methods, showcasing improvements in recognition accuracy ranging from 1.65% to 5.54%. These findings underscore the efficacy of our hybrid approach in advancing the field of human activity recognition, substantiating its superiority over both conventional baselines and advanced methodologies.

The comparative results of the PAMAP2 dataset show that the hybrid approach demonstrates certain advantages, with its recognition performance surpassing that of other state-of-the-art methods except for ConvAE-LSTM. Even when compared to ConvAE-LSTM, the hybrid approach lags behind in recognition accuracy by only 0.39%. However, the hybrid approach boasts lower model parameters and complexity, Our method reduces the overall parameter count by 19–21%, making it more advantageous for deployment on mobile devices. In summary, our human body pose model excels at extracting posture angles of various body parts to the fullest extent. By efficiently utilizing sensor data information, GAM-MLP effectively distinguishes and relates the differences and connections between actions when focusing on intra-group features. This process leads to precise recognition outcomes.

5.3. Recognition of Common Human Movements

Across the various human activity categories, the most prevalent ones include walking, standing, climbing upstairs, descending stairs, and lying down. In most scenarios, recognition often revolves around these primary actions. Therefore, we proceeded to extract these six key movement categories from both the MultiportGAM dataset and the PAMAP2 dataset. This focused evaluation allows us to assess our hybrid human pose recognition method's classification performance specifically on these commonly encountered activities.

The experimental outcomes, illustrated in Figures 12 and 13, highlight how the hybrid human pose recognition method performs when tasked with classifying these frequently observed actions within the MultiportGAM dataset and the PAMAP2 dataset.

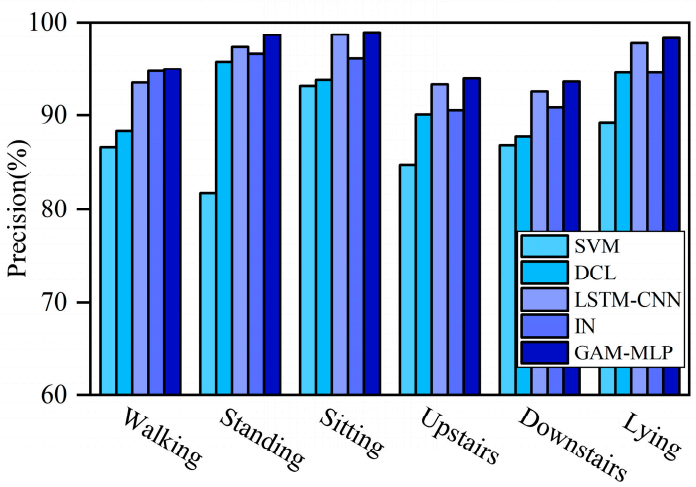


Figure 12. Classification performance of six commonly used actions on the MultiportGAM dataset.

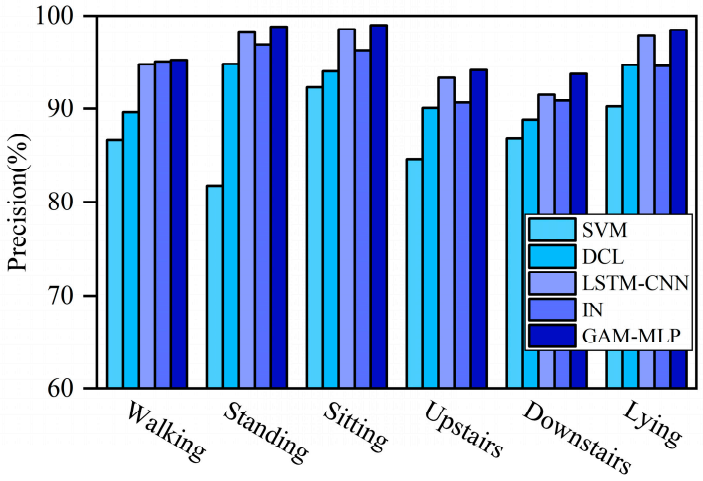


Figure 13. Classification performance of six commonly used actions on the PAMAP2 dataset.

It is evident that due to variations in the number and placement of sensors, the performance of GAM-MLP varies across different datasets. Comparing the two datasets, unlike the PAMAP2 dataset, the MultiportGAM dataset collects accelerometer, magnetometer, and gyroscope data from the chest and limbs, using two additional IMUs. In terms of daily behavior, the MultiportGAM has a longer duration, resulting in higher recognition accuracy.

On the MultiportGAM dataset, our GAM-MLP model achieves an average accuracy of 97.45% across the six common daily activities. This represents a 1.04–10.8% enhancement over other models. Notably, the highest accuracy is achieved in recognizing the "sitting down" action, reaching an accuracy of 98.94%. Even in the challenging "descending stairs" action, the model reaches an impressive precision of 93.67%. Similarly, on the PAMAP2 dataset, the "sitting down" action boasts the highest recognition rate of 99.02%, while the "descending stairs" action presents the lowest recognition rate at 93.74%. The average accuracy improves by 1.59–11.06% across the board. Through the combination of the multi-layer perceptron and group attention module, we achieve the information exchange and fusion between different features of the same group. In the training process, the neural network automatically learns and optimizes the weight and bias through the back-propagation

algorithm, so that the neurons can better capture the correlation information between different features, in order to improve the accuracy and reliability of the classification.

5.4. GAM Ablation Comparison

The experiment to evaluate the optimization effect of GAM was conducted on the MultiportGAM dataset and the PAMAP2 dataset, with GAM modules removed to compare the recognition performance of human movements. The comparison results on the MultiportGAM dataset are shown in Figure 14a, while the comparison results on the PAMAP2 dataset can be observed in Figure 14b.

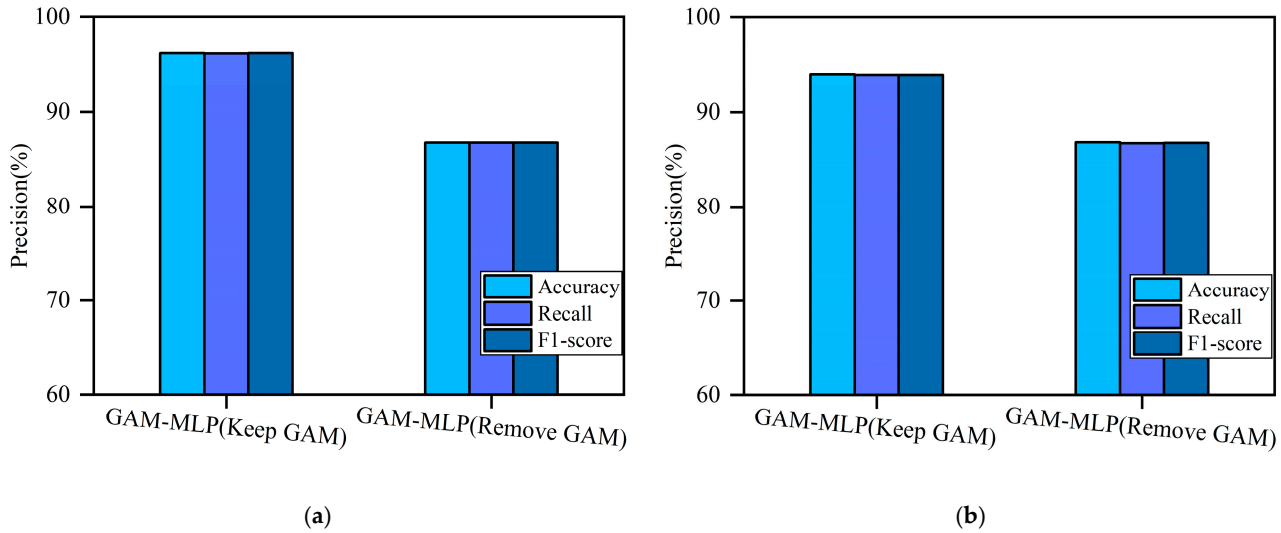


Figure 14. (**a**) Comparison of ablation experiments with preservation and removal of GAM on the MultiportGAM dataset. (**b**) Comparison of ablation experiments with preservation and removal of GAM on the PAMAP2 dataset.

The experimental results showed that on the MultiportGAM dataset, compared to the original structure without adding GAM, GAM achieved an accuracy improvement of 9.38%, a recall improvement of 9.37%, and an F1-score improvement of 9.37%. On the PAMAP2 dataset, these three indicators increased by 7.15%, 7.17%, and 7.16%, respectively.

To assess the generalization capability of the GAM module, we conducted a comparative analysis with other attention modules. These alternative attention modules were integrated into the residual blocks within the framework for experimentation, and the attention modules we compared included SE and CBAM. The MultiportGAM dataset and PAMAP2 dataset were employed as the dataset for our experiments. The experimental results are presented in Tables 5 and 6.

Table 5. Comparison of different methods for the MultiportGAM dataset.

Attention Module	Accuracy (%)	Recall (%)	F1-Score (%)
GAM-MLP	96.13	96.12	96.13
SE-MLP	92.74	92.13	92.48
CBAM-MLP	91.11	92.02	92.08

Table 6. Comparison of different methods for the PAMAP2 dataset.

Attention Module	Accuracy (%)	Recall (%)	F1-Score (%)
GAM-MLP	93.96	93.89	93.91
SE-MLP	89.34	88.75	89.16
CBAM-MLP	88.75	89.13	88.96

It can be observed that compared to GAM, the application of CE and CBAM, two general attention mechanisms, in the hybrid method, results in an accuracy improvement of about 3% compared to the proposed framework without the application of attention mechanisms. However, there is still a gap of 3.4–5% compared to GAM-MLP, indicating that GAM has a higher adaptability to the proposed framework. This stems from the manual grouping module of GAM, which reduces errors between specified groups while adapting to the active features of the dataset.

Thanks to GAM's ability to extract temporal features of sensor data and dynamically adjust the weights of accelerometer, gyroscope, and magnetometer data, GAM-MLP can better capture the associations between multiple sensors in different time periods, thereby improving the accuracy of action recognition.

5.5. Identification of 19 Types of Diverse Actions

In the recognition of 19 activities such as walking, jumping, and shooting baskets, GAM-MLP demonstrates exceptional capabilities due to the extraction of spatiotemporal features from raw sensor data that has been processed for activity construction. The model achieves an impressive average recognition accuracy of 96.13%. Notably, in the 3rd (Sitting), 4th (Lying), 15th (Cycling), and 19th (Rowing) activity categories, characterized by distinct features, the model achieves nearly complete recognition. While accuracy may slightly decrease in other activity categories, it still maintains a high precision of 93.54%. The confusion matrix for 19 types of action recognition is shown in Figure 15.

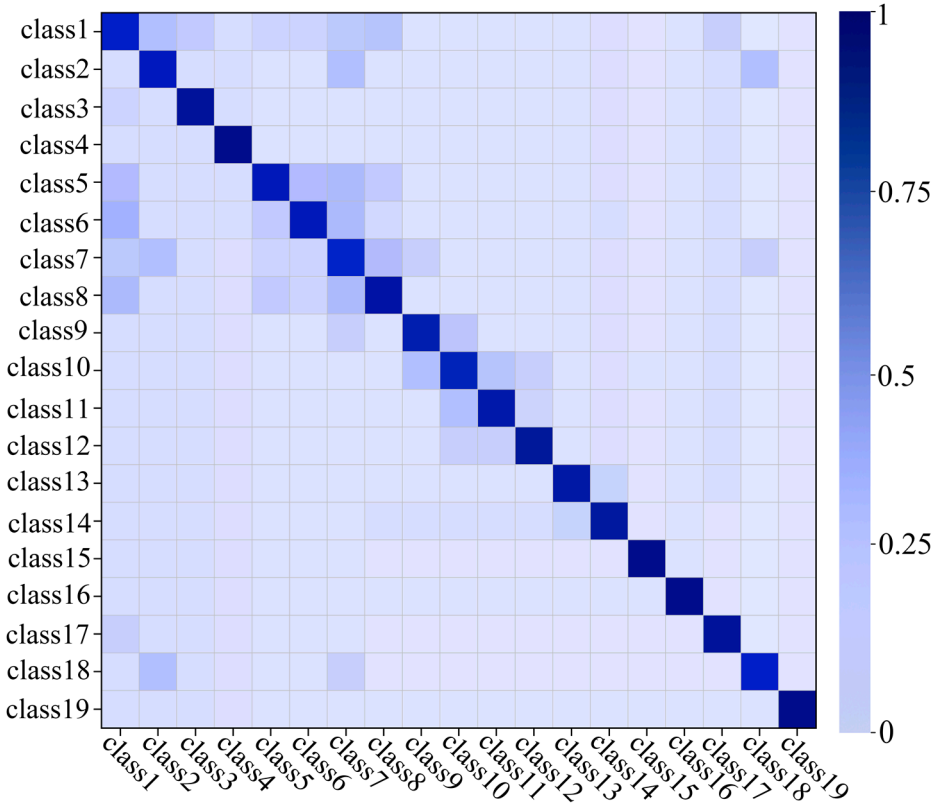


Figure 15. The confusion matrix for 19 types of action recognition on the MultiportGAM dataset.

In the recognition process of 19 types of activities, we used the attention mechanism to dynamically adjust the focus of the neural network, so as to better capture the key information and improve the accuracy of action recognition. This mechanism enables the neural network to adaptively focus on the important feature information, thereby avoiding the information redundancy and unnecessary computational overhead caused by the independent learning of the weight for each neuron in the conventional neural network.

6. Conclusions and Future Work

This paper offered a hybrid HAR approach using an MLP neural network and Euler angle extraction based on IMU sensors. In the method, by employing accelerometers, we precisely capture the acceleration of the torso, left arm, right arm, left leg, and right leg. This meticulous data collection approach is supplemented by gyroscopes, which monitor the angular velocity of various bodily components, and magnetometers, which gauge the direction and strength of magnetic fields. Refining joint placement through compensation of magnetometer and gyroscope data enhances the accuracy of human activity identification. Experimental findings, derived from tests conducted on widely used datasets, underscore the superiority of our proposed GAM-MLP model over existing deep learning-based models. On the PAMAP2 dataset and the MultiportGAM dataset, accuracy rates of 93.96% and 96.13% were achieved, respectively. Particularly noteworthy is the 97.45% accuracy achieved in recognizing six common daily activities within the MultiportGAM dataset. Compared to traditional methods and some advanced techniques, an improvement of 1.04% to 10.8% in accuracy was attained. Nevertheless, there is still room for improvement in GAM-MLP in identifying confusing actions

Looking forward, our future endeavors are geared towards exploring the realm of online deep learning models and expanding the horizons of both human activity recognition (HAR) and GAM-MLP to adeptly handle vast volumes of multivariate data. Specifically, in the future, additional inference modules can be considered in the GAM-MLP network to accelerate recognition speed and accuracy. When dealing with multivariate data, increasing the depth of the network appropriately may improve the recognition effect. Furthermore, the attention mechanism is intricately connected to the challenges posed by big data. Big data often exhibits diversity and variability, and attention mechanisms can adapt to various data distributions and characteristics, thereby enabling models to better accommodate shifts in the data. We are committed to enhancing the robustness of attention mechanisms in HAR recognition. They can assist models in focusing on crucial information, suppressing noise and outliers, and dynamically adjusting weights based on changing data, thereby maintaining model stability within the context of big data.

Author Contributions: Conceptualization, L.Y. and Y.M.; methodology, H.G.; software, Y.M.; validation, Y.Y. and X.H.; formal analysis, Y.H.; investigation, L.Y.; resources, H.G.; data curation, L.Y.; writing—original draft preparation, Y.M.; writing—review and editing, X.H. and Y.Y.; visualization, H.G.; supervision, Y.H.; project administration, X.H.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Zhejiang Province (No. LGG21F010005) and the National Natural Science Foundation of China under grant No. 61602137 and 61702144.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Written informed consent was obtained from the volunteer(s) to publish this paper.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xiao, Z.; Fu, M.; Yi, Y.; Lv, N. 3D Human Postures Recognition Using Kinect. In Proceedings of the 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, Nanchang, China, 26–27 August 2012; pp. 344–347.
2. Jaén-Vargas, M.; Reyes Leiva, K.; Fernandes, F.; Gonçalves, S.B.; Tavares Silva, M.; Lopes, D.S.; Serrano Olmedo, J. A Deep Learning Approach to Recognize Human Activity Using Inertial Sensors and Motion Capture Systems. In *Fuzzy Systems and Data Mining VII*; IOS Press: Amsterdam, The Netherlands, 2021.

3. Forsman, M.; Fan, X.; Rhen, I.M.; Lind, C.M. Mind the gap—Development of conversion models between accelerometer- and IMU-based measurements of arm and trunk postures and movements in warehouse work. *Appl. Ergon.* **2022**, *105*, 103841. [CrossRef] [PubMed]
4. Withanage, K.I.; Lee, I.; Brinkworth, R.; Mackintosh, S.; Thewlis, D. Fall Recovery Subactivity Recognition With RGB-D Cameras. *IEEE Trans. Ind. Inform.* **2016**, *12*, 2312–2320. [CrossRef]
5. Hoang, M.L.; Pietrosanto, A. Yaw/Heading optimization by drift elimination on MEMS gyroscope. *Sens. Actuators A Phys.* **2021**, *325*, 112691. [CrossRef]
6. Ito, C.; Cao, X.; Shuzo, M.; Maeda, E. Application of CNN for Human Activity Recognition with FFT Spectrogram of Acceleration and Gyro Sensors. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; pp. 1503–1510.
7. Bayraktar, E.; Yigit, C.B.; Boyraz, P. Object manipulation with a variable-stiffness robotic mechanism using deep neural networks for visual semantics and load estimation. *Neural Comput. Appl.* **2020**, *32*, 9029–9045. [CrossRef]
8. Yigit, C.B.; Bayraktar, E.; Boyraz, P. Low-cost variable stiffness joint design using translational variable radius pulleys. *Mech. Mach. Theory* **2018**, *130*, 203–219. [CrossRef]
9. Yigit, C.B.; Bayraktar, E.; Kaya, O.; Boyraz, P. External Force/Torque Estimation With Only Position Sensors for Antagonistic VSAs. *IEEE Trans. Robot.* **2021**, *37*, 675–682. [CrossRef]
10. Meng, Z.Z.; Zhang, M.X.; Guo, C.X.; Fan, Q.R.; Zhang, H.; Gao, N.; Zhang, Z.H. Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis. *Electronics* **2020**, *9*, 19. [CrossRef]
11. Anazco, E.V.; Lopez, P.R.; Park, H.; Park, N.; Kim, T.S. Human Activities Recognition with a Single Writs IMU via a Variational Autoencoder and Android Deep Recurrent Neural Nets. *Comput. Sci. Inf. Syst.* **2020**, *17*, 581–597.
12. Abdelhafiz, M.H.; Awad, M.I.; Sadek, A.; Tolbah, F. Sensor positioning for a human activity recognition system using a double layer classifier. *Proc. Inst. Mech. Eng. Part H J. Eng. Med.* **2021**, *236*, 248–258. [CrossRef]
13. Rivera, P.; Valarezo, E.; Kim, T.S. An Integrated ARMA-Based Deep Autoencoder and GRU Classifier System for Enhanced Recognition of Daily Hand Activities. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 19. [CrossRef]
14. Hashim, B.A.M.; Amutha, R. Deep transfer learning based human activity recognition by transforming IMU data to image domain using novel activity image creation method. *J. Intell. Fuzzy Syst.* **2022**, *43*, 2883–2890. [CrossRef]
15. Tahir, S.; Dogar, A.B.; Fatima, R.; Yasin, A.; Shafiq, M.; Khan, J.A.; Assam, M.; Mohamed, A.; Attia, E.A. Stochastic Recognition of Human Physical Activities via Augmented Feature Descriptors and Random Forest Model. *Sensors* **2022**, *22*, 20. [CrossRef] [PubMed]
16. Chakraborty, A.; Mukherjee, N. A deep-CNN based low-cost, multi-modal sensing system for efficient walking activity identification. *Multimed. Tools Appl.* **2023**, *82*, 16741–16766. [CrossRef]
17. Salem, Z.; Weiss, A.P. Improved Spatiotemporal Framework for Human Activity Recognition in Smart Environment. *Sensors* **2023**, *23*, 24. [CrossRef] [PubMed]
18. Fan, Y.; Jin, H.; Ge, Y.; Wang, N. Wearable Motion Attitude Detection and Data Analysis Based on Internet of Things. *IEEE Access* **2020**, *8*, 1327–1338. [CrossRef]
19. Wang, N.; Huang, J.; Yue, F.; Zhang, X. Attitude Algorithm and Calculation of Limb Length Based on Motion Capture Data. In Proceedings of the 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China, 27–31 December 2021; pp. 1004–1009.
20. Heng, X.; Wang, Z.; Wang, J. Human activity recognition based on transformed accelerometer data from a mobile phone. *Int. J. Commun. Syst.* **2016**, *29*, 1981–1991. [CrossRef]
21. Xiao, X.; Zarar, S. In A Wearable System for Articulated Human Pose Tracking under Uncertainty of Sensor Placement. In Proceedings of the 7th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, BIOROB, Enschede, The Netherlands, 26–29 August 2018; IEEE Computer Society: Enschede, The Netherlands, 2018; pp. 1144–1150.
22. Cui, C.; Li, J.; Du, D.; Wang, H.; Tu, P.; Cao, T. The Method of Dance Movement Segmentation and Labanotation Generation Based on Rhythm. *IEEE Access* **2021**, *9*, 31213–31224. [CrossRef]
23. Shenoy, P.; Sompur, V.; Skm, V. Methods for Measurement and Analysis of Full Hand Angular Kinematics Using Electromagnetic Tracking Sensors. *IEEE Access* **2022**, *10*, 42673–42689. [CrossRef]
24. Aasha, M.; Sivaranjani, S.; Sivakumari, S. An Effective reduction of Gait Recognition Time by using Gender Classification. In Proceedings of the International Conference on Advances in Information Communication Technology & Computing—AICTC '16, Bikaner, India, 12–13 August 2016; pp. 1–6.
25. Chen, Y.; Tu, Z.; Kang, D.; Chen, R.; Bao, L.; Zhang, Z.; Yuan, J. Joint Hand-Object 3D Reconstruction From a Single Image With Cross-Branch Feature Fusion. *IEEE Trans. Image Process.* **2021**, *30*, 4008–4021. [CrossRef]
26. Cui, Y.; Li, X.; Wang, Y.; Yuan, W.; Cheng, X.; Samiei, M. MLP-TLBO: Combining Multi-Layer Perceptron Neural Network and Teaching-Learning-Based Optimization for Breast Cancer Detection. *Cybern. Syst.* **2022**, *53*, 1–28. [CrossRef]
27. Faundez-Zanuy, M.; Ferrer-Ballester, M.A.; Travieso-González, C.M.; Espinosa-Duro, V. Hand Geometry Based Recognition with a MLP Classifier. In *Advances in Biometrics*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 721–727.
28. Yuan, X.; Liang, N.; Fu, W.; Wang, Q.; Zhang, Y.; Cao, J.; Liu, H.; Liu, K.; Huang, Y.; Ren, X. A Wearable Gesture Recognition System With Ultrahigh Accuracy and Robustness Enabled by the Synergy of Multiple Fabric Sensing Devices. *IEEE Sens. J.* **2023**, *23*, 10950–10958. [CrossRef]

29. Anwar, I.N.; Daud, K.; Samat, A.A.A.; Soh, Z.H.C.; Omar, A.M.S.; Ahmad, F. Implementation of Levenberg-Marquardt Based Multilayer Perceptron (MLP) for Detection and Classification of Power Quality Disturbances. In Proceedings of the 022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 21–22 October 2022; pp. 63–68.
30. Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; Moreno-Noguer, F. Back to MLP: A Simple Baseline for Human Motion Prediction. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 4798–4808.
31. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* **2021**, *102*, 107101. [CrossRef]
32. Vasyltsov, I.; Chang, W. Efficient softmax approximation for deep neural networks with attention mechanism. *arXiv* **2021**, arXiv:2111.10770.
33. Ordonez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, *16*, 115. [CrossRef] [PubMed]
34. Xu, C.; Chai, D.; He, J.; Zhang, X.T.; Duan, S.H. InnoHAR: A Deep Neural Network for Complex Human Activity Recognition. *IEEE Access* **2019**, *7*, 9893–9902. [CrossRef]
35. Xia, K.; Huang, J.; Wang, H. LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access* **2020**, *8*, 56855–56866. [CrossRef]
36. Lv, T.; Wang, X.; Jin, L.; Xiao, Y.; Song, M. Margin-Based Deep Learning Networks for Human Activity Recognition. *Sensors* **2020**, *20*, 1871. [CrossRef]
37. Wan, S.; Qi, L.; Xu, X.; Tong, C.; Gu, Z. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. *Mob. Netw. Appl.* **2020**, *25*, 743–755. [CrossRef]
38. Huang, W.; Zhang, L.; Wu, H.; Min, F.; Song, A. Channel-Equalization-HAR: A Light-weight Convolutional Neural Network for Wearable Sensor Based Human Activity Recognition. *IEEE Trans. Mob. Comput.* **2023**, *22*, 5064–5077. [CrossRef]
39. Tang, Y.; Zhang, L.; Min, F.; He, J. Multiscale Deep Feature Learning for Human Activity Recognition Using Wearable Sensors. *IEEE Trans. Ind. Electron.* **2023**, *70*, 2106–2116. [CrossRef]
40. Thakur, D.; Biswas, S.; Ho, E.S.L.; Chattopadhyay, S. ConvAE-LSTM: Convolutional Autoencoder Long Short-Term Memory Network for Smartphone-Based Human Activity Recognition. *IEEE Access* **2022**, *10*, 4137–4156. [CrossRef]