Transformer-Enabled Smartphone System for Intelligent Physical Activity Monitoring

Leping Zhang[1], Fengjiao Jiang[2]*, Guopeng Jia[3], Yue Wang[4]
Personnel Department, Shanghai Vocational College of Agriculture and Forestry, Shanghai, China[1]
Department of Intelligent Agricultural Engineering, Shanghai Vocational College of Agriculture and Forestry, Shanghai, China[2, 3, 4]

***Abstract*—This study addresses the prevalent decline in physical activity among university students in the contemporary information society, proposing an innovative deep learning-based framework for intelligent physical activity recognition. Central to this framework is the comprehensive utilization of high-precision Inertial Measurement Units (IMUs) integrated within smartphones, encompassing triaxial accelerometers, gyroscopes, and magnetometers, enabling multi-dimensional, real-time capture of students' daily activity postures. For algorithmic design, this research transcends traditional limitations by adopting the more advanced Transformer architecture as its core classifier. Through the distinct self-attention mechanism inherent to this architecture, the proposed method efficiently and precisely extracts critical spatiotemporal features from vast sensor data, thereby achieving accurate identification and classification of various physical activities, such as walking, running, and climbing stairs. Rigorous evaluation results demonstrate significant advantages in key performance metrics, including recognition accuracy, when compared to conventional recurrent neural networks (e.g., Long Short-Term Memory networks, Recurrent Neural Networks) and classic machine learning algorithms (e.g., Random Forest), with a validation accuracy reaching 93.97%. This forward-looking research outcome not only provides a reliable and efficient technological means for monitoring the physical activity status of university students but also establishes a robust data foundation for the future development and implementation of targeted health intervention measures.**

Keywords*—Activity recognition; smartphone; transformer architecture; inertial measurement units

I. Introduction

In recent years, the widespread adoption of smartphones and rapid advancements in mobile computing technology have positioned sensor-based Human Activity Recognition as a focal point in research fields such as health monitoring, intelligent assistance, and sports analytics [1]. Traditional methods for monitoring physical activity, such as questionnaires, field-based equipment assessments, or manual observation, suffer from limitations including subjectivity, high costs, and poor compliance, making them inadequate for large-scale, long-term, and precise monitoring. Consequently, there is an urgent need for intelligent, convenient, and efficient physical activity recognition technologies to address the shortcomings of conventional methods.

The perception of physical activities forms the cornerstone of intelligent recognition. Currently, the primary approaches to physical activity perception encompass time-series data-based sensing and image-based visual sensing [2]. The use of diverse sensor modalities for activity detection and classification has emerged as a transformative technology for real-time and autonomous monitoring in areas such as behavioral analysis in smart home environments, assisted living, daily activity monitoring, elderly care, rehabilitation, entertainment, and security surveillance. Wearable devices, smartphones [3], and ambient sensing equipment are equipped with an array of sensors, including accelerometers, gyroscopes, magnetometers, heart rate monitors, pressure sensors, and compact cameras, to facilitate activity recognition and monitoring [4]. These sensor data undergo preprocessing to extract feature sets, such as time-domain, frequency-domain, or wavelet transform features, which are then processed using machine learning algorithms to classify and continuously monitor human activities.

The perception of time-series data, such as acceleration signals, provides a convenient approach for activity recognition. By integrating these data with intelligent algorithms, accurate identification of human activities can be achieved. Yin et al. developed a two-stage method for detecting anomalous activities using wireless body sensors [5]. In the first stage, they trained a one-class Support Vector Machine (SVM) on common normal activities to filter out those highly likely to be normal. Subsequently, they used Kernel Nonlinear Regression (KNLR) to derive an anomaly activity model from the general normal model, reducing false positives in an unsupervised manner. Saeed et al. proposed a novel self-supervised technique for feature learning from sensor data, which did not require semantic labels (i.e., activity categories) [6]. They trained a multi-task temporal convolutional network to recognize transformations applied to input signals. By leveraging these transformations, they demonstrated that a simple binary auxiliary task could generate robust supervisory signals, extracting valuable features for downstream tasks. Janidarmian et al. performed an extensive analysis of feature representations and classification techniques for activity recognition, comparing 293 classifiers in the most comprehensive study to date [7]. They applied Principal Component Analysis (PCA) to reduce feature vector dimensionality while preserving essential information. Iloga et al. observed that deep learning techniques addressed some limitations but required significant computational resources and produced feature vectors with limited interpretability [8]. To address these challenges, they developed a Human Activity Recognition (HAR) technique based on Hidden Markov Models (HMMs).

*Corresponding Author.

Smartphones equipped with various sensors, such as accelerometers, gyroscopes, and magnetometers, can capture high-frequency data on posture, orientation, and velocity changes during human motion, generating continuous time-series data [9]. These data accurately capture unique patterns of periodic or non-periodic movements, such as running, jumping, and walking [10]. For instance, accelerometers record linear motion, while gyroscopes detect rotational motion, together providing a comprehensive representation of the body's dynamics in three-dimensional space [11]. Through preprocessing, feature extraction, and pattern recognition of sensor data, different physical activities can be effectively distinguished [12]. Compared to other wearable devices, smartphones, as indispensable tools in university students' daily lives, offer widespread availability and portability, making them an ideal platform for time-series data collection in physical activity sensing [13]. Lara et al. developed the Centinela system [14], which integrated acceleration data with vital signs to achieve high-accuracy activity recognition. The system identified five activities: walking, running, sitting, standing, and descending. They evaluated eight classifiers and three time window sizes, achieving an overall accuracy of 95.7%. Micucci et al. created a novel acceleration sample dataset [15], collected using an Android smartphone designed for human activity recognition and fall detection. The samples were categorized into 17 fine-grained classes and grouped into two coarse-grained classes: one comprising 9 types of Activities of Daily Living (ADL) and another including 8 types of falls. Mario proposed a novel mechanism for detecting specific activities using data from a single triaxial accelerometer [16]. They employed convolutional neural networks to automatically extract the most relevant features to characterize acceleration patterns, enabling cross-activity recognition.

With significant advancements in deep learning for image processing, intelligent recognition of human activities based on image data has seen rapid development. Niu et al. developed a framework for human activity detection and recognition in outdoor video surveillance applications [17]. They introduced an efficient activity representation method that identified distinct interaction patterns among groups based on simple statistical data from tracking trajectories, without requiring complex Markov chains, Hidden Markov Models (HMMs), or Coupled Hidden Markov Models (CHMMs). Sung et al. utilized RGBD sensors (Microsoft Kinect) as input devices and computed a set of features based on human posture, motion, and image and point cloud data [18]. They designed an algorithm based on a hierarchical Maximum Entropy Markov Model (MEMM), treating individual activities as compositions of sub-activities, and employed dynamic programming to infer a two-layer graph structure. Ni et al. proposed a novel framework for complex activity recognition and localization, effectively integrating information from grayscale and depth image channels across multiple layers of the video processing pipeline [19]. Koppula et al. modeled complex spatiotemporal relationships (termed affordances) between human postures and objects using Conditional Random Fields (CRFs), inferring multiple possible graph structures and approximating graphs with additive features for efficient dynamic programming [20]. Albanese et al. developed a computational framework for human activity representation based on Petri Nets [21]. They introduced the PPN-MPS algorithm for the first question and proposed two algorithms (naive PPN-MPA and PPN-MPA) for the second. Rodrigues et al. developed a multi-timescale model to capture temporal dynamics across different time scales, predicting future and past states for a given input posture trajectory [22].

In the domain of physical activity recognition, deep learning techniques have demonstrated exceptional capabilities in feature extraction and pattern recognition, establishing them as the most advanced and effective algorithmic paradigm. For time-series data collected by smartphones, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants, such as Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), have been widely applied to physical activity recognition [23]. Additionally, models incorporating attention mechanisms further enhance the ability to capture critical motion features [24]. For image data, CNNs remain the dominant algorithm for physical activity recognition. Transfer learning based on pretrained ImageNet models significantly improves performance in image-based activity recognition tasks. By leveraging posture information derived from human skeleton keypoint detection, combined with algorithms like Graph Convolutional Networks (GCNs), complex physical activities can be analyzed more effectively through posture changes and action sequences. Furthermore, Transformer models have shown significant potential in processing image sequences and capturing global features, opening new research directions for image-based physical activity recognition [25].

Transformer technology continues to demonstrate significant potential in processing time-series and tabular data. To this end, this study proposes a novel method for intelligent recognition of physical activities in students using Transformer-based models and smartphone sensors. The method leverages high-precision inertial sensors embedded in smartphones, such as accelerometers and gyroscopes, to collect motion data. Through the Transformer neural network, feature extraction and classification are performed to automatically identify various physical activities, including walking, running, and climbing stairs. Compared to traditional time-series models such as Long Short-Term Memory (LSTM) networks, vanilla Recurrent Neural Networks (RNNs), and machine learning methods like Random Forest, the proposed method demonstrates a significant advantage in recognition accuracy. In terms of algorithmic architecture, this study pioneers the application of Transformer models in the domain of university students' daily activity recognition. Leveraging its distinctive self-attention mechanism, the proposed framework achieves efficient and precise extraction of crucial spatiotemporal features from massive sensor datasets. Regarding practical applications, the research not only develops a reliable and efficient monitoring system for collegiate physical activities but more significantly incorporates an innovative class-balancing mechanism that effectively mitigates recognition performance degradation caused by the data imbalance issue.

Section II provides a detailed introduction to the overall framework and the intelligent recognition algorithm of the proposed method. Section III conducts an in-depth analysis of

the utilized data, and Section IV evaluates the recognition performance of the proposed method in comparison with other approaches. Section V presents a discussion and analysis. Finally, Section VI summarizes the findings of this study.

II. Portable Perception and Intelligent Recognition of Human Activities

To accurately and comprehensively monitor and assess the types and states of physical activities among university students, this study innovatively proposes a portable monitoring method based on smartphone technology and deep learning. Its overall framework, clearly illustrated in Fig. 1, delineates the entire data flow and processing pipeline. This method capitalizes on the pervasive and ubiquitous nature of modern smartphones, specifically leveraging their integrated high-precision inertial sensors, such as triaxial accelerometers and triaxial gyroscopes. These sensors enable the portable sensing of human activities, capturing detailed motion data with remarkable accuracy, thereby establishing a robust foundation for subsequent analytical procedures. In the data processing phase, the raw sensor signals undergo a series of meticulous preprocessing steps. Specifically, the data is first subjected to precise segmentation in both the time and frequency domains, dividing it into fixed-length windows to ensure that each window encapsulates complete activity information. Subsequently, these segmented data are normalized to eliminate scale discrepancies between different sensors or individuals, ensuring data consistency. Following normalization, a rich set of time-domain and frequency-domain features is extracted from the signals, designed to capture the unique patterns of distinct physical activities. These extracted features collectively form a structured tabular dataset, where each sample comprises a carefully selected array of diverse features, preparing the data for input into the machine learning model.

Building upon this foundation, this research innovatively designs and proposes an optimized Transformer network model specifically tailored for this tabular data. This model integrates advanced deep learning techniques, including a unique mapping projection mechanism and a sophisticated positional encoding strategy. These techniques enable the Transformer network to effectively process the non-sequential structure inherent to tabular data and capture complex relationships and potential temporal dependencies among features (even though the data has been tabularized). Through this powerful model, we achieve highly accurate classification of various human activities, thereby robustly supporting the continuous monitoring and assessment of university students' activity types and states. This portable, efficient, and precise monitoring method promises to offer a revolutionary tool for managing the physical health of university students.

The physical activity recognition model proposed in this study leverages the robust modeling capabilities of the Transformer architecture combined with the characteristics of multidimensional feature data. Originally designed for natural language processing tasks, the Transformer model has gained widespread attention due to its superior sequence modeling and global information capture mechanisms. The model processes input tabular data through four core components—input projection, positional encoding, Transformer encoder, and classification head—to capture complex dependencies among features and generate probabilistic predictions for activity categories. Compared to traditional vanilla Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), the Transformer architecture employs a self-attention mechanism to process tabular data in parallel, significantly enhancing its ability to model multidimensional features while mitigating issues such as vanishing gradients. The input projection module maps raw tabular data features into a high-dimensional model space, improving feature representation. The positional encoding module introduces positional information to the tabular data, addressing the Transformer's inherent lack of sequence awareness. The Transformer encoder, utilizing multiple layers of multi-head self-attention mechanisms and feed-forward networks, captures intricate dependencies within the tabular data. The classification head pools the encoded sequence features and maps them to the category space, producing the final activity predictions. Each component incorporates recent advances in deep learning, ensuring the model's efficiency and robustness in handling time-series data. The following sections provide a detailed description of each module.

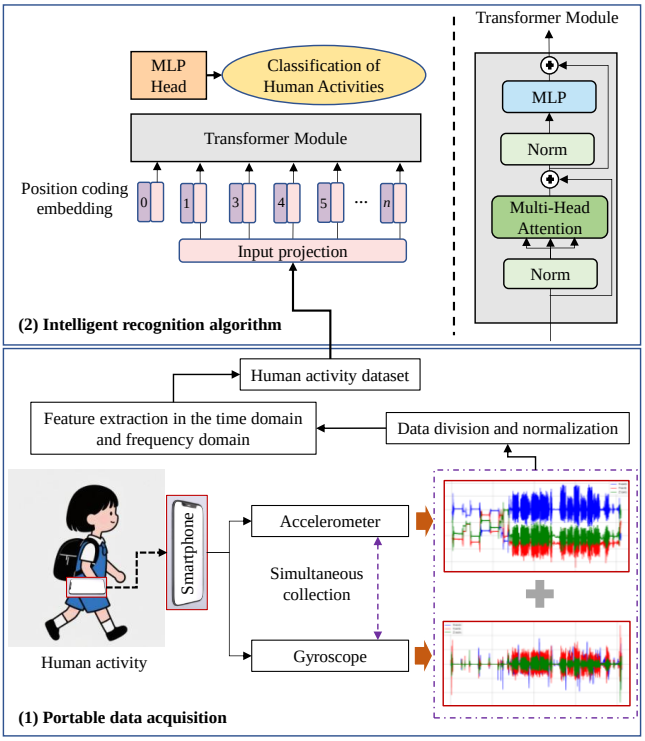


Fig. 1. Overall framework of the proposed method.

A. Projection of the Input Data

The input projection module aims to map raw tabular data features ($\mathbf{X} \in \mathrm{R}^{B \times D}$) to the model's internal dimension (d_{model}) to meet the input requirements of the Transformer encoder while enhancing feature representation through nonlinear transformations. The projection process encompasses linear transformation, layer normalization, activation functions, and dropout operations, defined as follows:

$$\mathbf{Z}_0 = \text{Dropout}(\text{ReLU}(\text{LayerNorm}(\mathbf{W}_{\text{in}}\mathbf{X} + \mathbf{b}_{\text{in}}))) \quad (1)$$

where, $\mathbf{W}_{\text{in}} \in \mathrm{R}^{D \times d_{\text{model}}}$, $\mathbf{b}_{\text{in}} \in \mathrm{R}^{d_{\text{model}}}$ represent the learnable parameters of the linear transformation. Layer normalization (LayerNorm) is defined as follows:

$$\text{LayerNorm}(\mathbf{x}) = \gamma \cdot \frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 + \grave{o}}} + \beta \quad (2)$$

where, μ and σ^2 denote the mean and variance, respectively, $\gamma, \beta \in \mathrm{R}^{d_{\text{model}}}$ are learnable parameters, and $\grave{o}$ is a small constant to prevent division by zero. The ReLU activation function introduces nonlinearity, enhancing the model's expressive capacity. Dropout randomly zeros elements with probability (p_{dropout}), mitigating overfitting. Layer normalization standardizes the feature distribution, alleviating internal covariate shift and improving training stability. Together, the ReLU activation function and dropout operations enhance the model's nonlinear modeling capability and generalization performance.

B. Position Coding Transformer

The Transformer architecture is inherently insensitive to the order of input tabular data, necessitating the explicit addition of positional information to preserve the sequential characteristics of the data. To retain the order information of tabular data features, this study employed sine and cosine functions to generate positional encodings, defined as:

$$\text{PE}(t, 2i) = \sin\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right), \quad \text{PE}(t, 2i+1) = \cos\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right) \quad (3)$$

where, $t \in \{0, 1, \ldots, \text{max\_len} - 1\}$ denotes the feature step, and $i \in \{0, 1, \ldots, (d_{\text{model}}/2) - 1\}$ represents the dimension index. max_len is the maximum feature length. The positional encoding is added to the projected input, followed by a dropout operation, defined as follows:

$$\mathbf{Z}_1 = \text{Dropout}(\mathbf{Z}_0 + \mathbf{PE}_{[:T]}) \quad (4)$$

where, $\mathbf{PE}_{[:T]} \in \mathrm{R}^{T \times d_{\text{model}}}$ denotes the positional encoding truncated to the first T rows. The sine and cosine functions generate periodic signals, enabling the model to distinguish between different features while maintaining smooth frequency variations. Fixed positional encodings, as opposed to learnable encodings, reduce the number of parameters and offer better generalization across tabular data of varying lengths. The dropout operation further enhances model robustness, preventing over-reliance on positional encodings.

C. Transformer Encoder

The Transformer encoder serves as the core component of the model, tasked with capturing complex dependencies among features in tabular data. It consists of L stacked encoder layers, each comprising a multi-head self-attention mechanism and a feed-forward neural network (FFN), augmented by residual connections and layer normalization. For the l-th layer, given the input ($\mathbf{Z}_{l-1} \in \mathrm{R}^{B \times T \times d_{\text{model}}}$), the multi-head self-attention is computed as follows:

$$\mathbf{A}_h = \text{Softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^{\cdot}}{\sqrt{d_k}}\right)\mathbf{V}_h \quad (5)$$

Here, ($\mathbf{Q}_h = \mathbf{Z}_{l-1}\mathbf{W}_{Q,h}, \mathbf{K}_h = \mathbf{Z}_{l-1}\mathbf{W}_{K,h}, \mathbf{V}_h = \mathbf{Z}_{l-1}\mathbf{W}_{V,h}$) represent the query, key, and value matrices, respectively, $\mathbf{W}_{Q,h}, \mathbf{W}_{K,h}, \mathbf{W}_{V,h} \in \mathrm{R}^{d_{\text{model}} \times d_k}; d_k = d_{\text{model}}/n_{\text{head}}; n_{\text{head}}$ denotes the number of attention heads, $\sqrt{d_k}$ is the scaling factor used to stabilize gradient computations, and $h \in \{1, \ldots, n_{\text{head}}\}$ represents the multi-attention head. The multi-head self-attention mechanism enables the model to focus on different parts of the sequence in parallel, capturing diverse feature dependency patterns. Residual connections preserve the original input information, mitigating gradient vanishing issues in deep networks. The feed-forward neural network (FFN) applies a nonlinear transformation independently to each time step, enhancing feature representation, and is defined as follows:

$$\mathbf{F} = \text{ReLU}(\mathbf{Z}_l^{'}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (6)$$

Here, $\mathbf{W}_1 \in \mathrm{R}^{d_{\text{model}} \times d_{\text{ff}}}, \mathbf{W}_2 \in \mathrm{R}^{d_{\text{ff}} \times d_{\text{model}}}, \mathbf{b}_1 \in \mathrm{R}^{d_{\text{ff}}}, \mathbf{b}_2 \in \mathrm{R}^{d_{\text{model}}}, d_{\text{ff}}$ denotes the dimension of the feed-forward neural network. The FFN enhances the model's nonlinear modeling capability, capturing complex data patterns. Residual connections and layer normalization further improve training stability, enabling the stacking of additional layers to increase the model's capacity.

D. Classification Head

The classification head employs a MultiLayer Perceptron (MLP) to map the pooled feature vectors to a probability distribution over activity categories, generating the final predictions for physical activities. Specifically, the classification head first applies a global pooling operation, such as mean pooling or max pooling, to the sequence features output by the Transformer encoder, extracting representative global features. These features are then fed into a multilayer perceptron consisting of multiple fully connected layers, each incorporating nonlinear activation functions (e.g., ReLU) and dropout operations to enhance the model's expressive power and generalization performance. Finally, a softmax function is applied to map the output to a probability distribution over the activity categories, enabling precise classification of physical activities.

$$\mathbf{Y} = \text{Softmax}(\text{Dropout}(\text{ReLU}(\text{LayerNorm}(\mathbf{Z}_{\text{pool}}\mathbf{W}_3 + \mathbf{b}_3)))\mathbf{W}_4 + \mathbf{b}_4) \quad (7)$$

Here, $\mathbf{W}_3 \in \mathrm{R}^{d_{\text{model}} \times d_{\text{ff}}}, \mathbf{W}_4 \in \mathrm{R}^{d_{\text{ff}} \times C}, \mathbf{b}_3 \in \mathrm{R}^{d_{\text{ff}}}, \mathbf{b}_4 \in \mathrm{R}^{C}$. The deep classification head enhances the model's classification capability through additional linear layers, layer normalization, and nonlinear activation functions. Dropout operations effectively mitigate overfitting, improving the model's generalization performance on test data. The softmax function

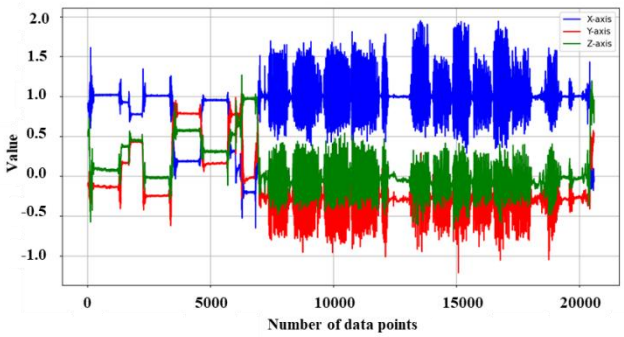
ensures that the output forms a valid probability distribution, making it suitable for multi-class classification tasks.

III. Analysis of Human Activity Perception Data

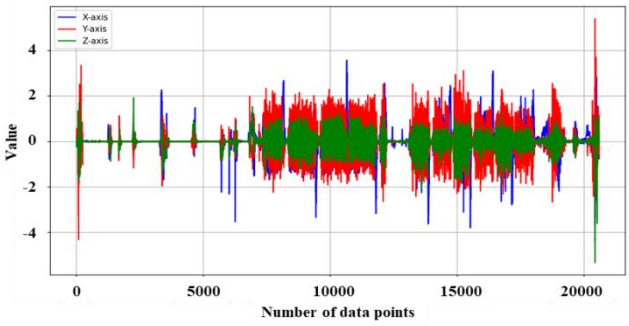
The proposed method facilitates efficient and accurate human activity recognition. The HAPT (Human Activity Recognition Using Smartphones) dataset [26], a widely adopted benchmark in HAR research, and available through the UCI Machine Learning Repository, captures diverse daily activities using smartphone-embedded inertial sensors. The dataset comprises time-series signals collected at 50 Hz from a waist-mounted smartphone, including: triaxial accelerometer data (X/Y/Z axes) quantifies linear acceleration patterns associated with body movements such as walking, sitting, or transitioning between postures. Triaxial gyroscope data (X/Y/Z axes) measures angular velocities to characterize rotational dynamics during activities like turning or limb motion. This multimodal sensor fusion provides complementary kinematic representations, enabling robust activity classification while maintaining temporal granularity critical for motion analysis. The standardized sampling protocol ensures consistent signal resolution across all recorded activities. The dataset for this study was collected from 30 volunteers aged between 19 and 48 years. During the experiments, a Samsung Galaxy S II smartphone was securely placed at the waist of each participant. This specific placement was chosen to optimize the capture of core trunk movements, thereby facilitating a more accurate recognition of whole-body activities. Participants were instructed to perform 12 predefined daily activities, each explicitly labeled. Fig. 2 illustrates the triaxial accelerometer and triaxial gyroscope signals recorded from one participant during an activity. The raw signals were preprocessed for noise reduction using a median filter and a third-order low-pass Butterworth filter with a cutoff frequency of 20 Hz. The clarity and distinctness of these signals demonstrate the smartphone's capability to accurately and effectively capture human motion data during activities. This provides compelling evidence for the feasibility of using smartphones as a robust tool for human activity recognition.

The complete dataset comprises a substantial volume of sensor readings. Following meticulous preprocessing and segmentation, these readings are transformed into a large collection of time-series samples. Data segmentation involves dividing continuous sensor data into fixed-length windows, typically 2.56 seconds (equivalent to 128 sampling points). Overlapping between windows is often employed to ensure each window comprehensively captures complete activity information. During the feature extraction phase, a diverse set of time-domain and frequency-domain features is extracted from the raw accelerometer and gyroscope signals for each 2.56-second window. These features are specifically designed to encapsulate the unique patterns associated with different human activities.

Following the crucial feature extraction step, each individual time window is successfully transformed into a high-dimensional feature vector. For instance, in the HAPT dataset, each such sample is represented by 561 carefully selected features, and these high-dimensional vectors subsequently serve as input data for training and inference with various machine learning models. To offer a more intuitive perspective, allowing for a clear understanding of the unique distribution patterns of these feature values across different activity types. Fig. 3 illustrates the 561 feature values for a single sample within a specific activity class. From the visual representation in this figure, it is evident that all 12 distinct human activity types, including walking, ascending stairs, descending stairs, sitting, standing, lying, and a range of complex transitional movements (such as standing-to-sitting, sitting-to-standing, sitting-to-lying, lying-to-sitting, standing-to-lying, and lying-to-standing), exhibit notably distinct feature curves. Specifically, when the human body is engaged in dynamic physical activities, the extracted feature values tend to be quantitatively higher. Conversely, when the body is in a static or non-moving state, such as sitting or lying, the corresponding feature values are generally observed to be at lower levels. A particularly interesting phenomenon is observed during standing posture, where the distribution of feature values demonstrates an intriguing duality: approximately half of the feature values display relatively high magnitudes, while the other half are comparatively smaller. Although these differences in feature values indeed provide a basis for distinguishing between various activity types and demonstrate a certain discriminatory capability, recognizing the inherent and significant inter-individual variability in behavioral characteristics, we acknowledge the limitations of existing methods. Therefore, future research urgently needs to develop more advanced and intelligent methods and algorithms to achieve more accurate and faster recognition and classification of human physical activity types, thereby better adapting to the complex and diverse patterns of human movement in real-world scenarios.

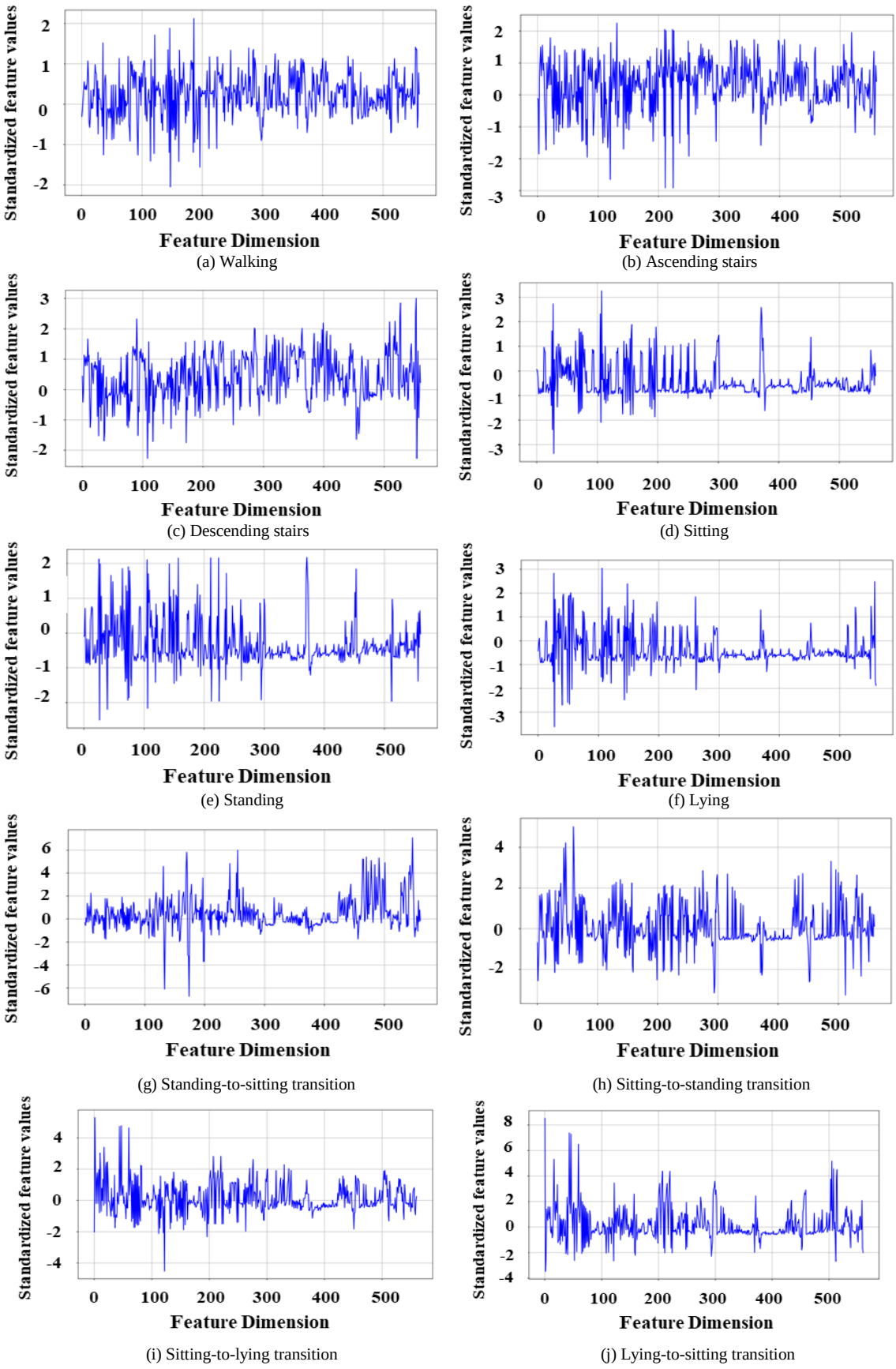


(a) Variation curves of the three-axis acceleration (m/s2).



(b) Variation curves of the three-axis gyroscope (rad/s).

Fig. 2. Two types of perception signal curves of a certain tester.



(a) Walking

(b) Ascending stairs

(c) Descending stairs

(d) Sitting

(e) Standing

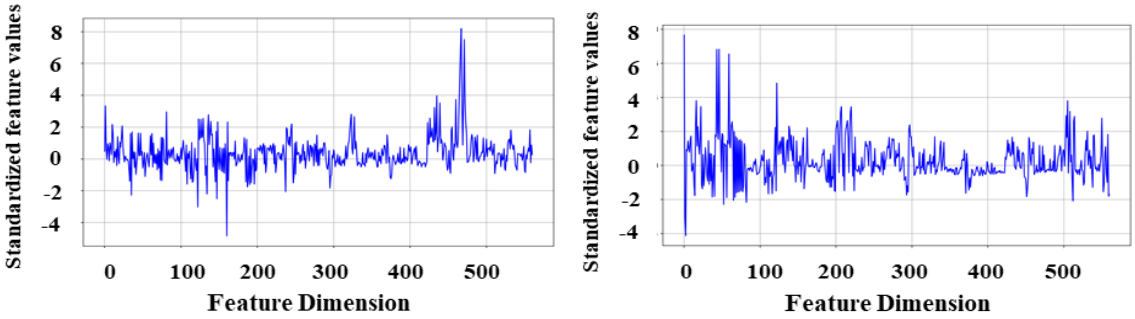
(f) Lying

(g) Standing-to-sitting transition

(h) Sitting-to-standing transition

(i) Sitting-to-lying transition

(j) Lying-to-sitting transition



(k) Standing-to-lying transition (l) Lying-to-standing transition

Fig. 3. Display of characteristic curves of different types of activities.

The class distribution of the dataset utilized in this study is presented in Table I. The training set comprises 7,767 samples, while the test set contains 3,153 samples, yielding a split ratio of approximately 2.46:1. This partitioning is considered reasonable and aligns with conventional machine learning data splitting standards, such as 70%-30% or 80%-20% distributions. Within the dataset, Classes 1 to 6 are identified as major categories, exhibiting substantial sample sizes. Specifically, each of these classes contains over 1,000 samples in the training set and more than 400 samples in the test set, collectively constituting the vast majority of the data. Conversely, Classes 7 to 12 are designated as minority categories, characterized by extremely sparse sample representation. For instance, the training set for minority classes contains no more than 90 samples each, and the test set no more than 50 samples each. Notably, Class 8 is the most underrepresented, with only 23 samples in the training set and 10 samples in the test set, indicating a severe class imbalance issue within the dataset.

To elaborate, the largest class, Class 5, contains 1,423 samples in the training set, which is 62 times larger than the smallest class, Class 8, with only 23 samples; a similar disparity is observed in the test set. Furthermore, the combined total of minority classes (Classes 7 to 12) accounts for only 4.5% of the total samples in the training set and 5.3% in the test set. This extreme imbalance implies that conventional evaluation metrics, such as overall accuracy, may be predominantly influenced by the majority classes, thereby potentially masking performance deficiencies on the minority classes. It is worth noting that the proportion of samples for each class is largely consistent between the training and test sets. For instance, Class 1 accounts for 15.8% of the training set and 15.7% of the test set. This consistency suggests that the data partitioning maintained distributional uniformity, effectively mitigating potential biases introduced by the split. Nevertheless, the dataset presents a small-class risk: certain minority classes, such as Classes 8, 10, and 12, possess extremely low sample counts in the test set (no more than 27 samples each), which could lead to significant volatility in evaluation results. Overall, the dataset exhibits a pronounced long-tail distribution, where the insufficient number of samples in minority classes may constrain the model's generalization capabilities.

TABLE I. ANALYSIS OF THE CLASS DISTRIBUTION OF THE DATASET

Class	Train dataset	Test dataset
1	1226	496
2	1073	462
3	987	420
4	1293	508
5	1423	556
6	1413	545
7	47	23
8	23	10
9	75	32
10	60	25
11	90	49
12	57	27
Total	7767	3153

IV. COMPARATIVE TESTING AND ANALYSIS

A. Model Training and Evaluation

In order to fully verify the accuracy and feasibility of the proposed method, human perception data is used to update the constructed recognition network. In addition, three other classic machine learning algorithms were used for comparison. The assessment results are shown in Table II. All evaluated models demonstrated excellent performance on the training set, clearly indicating their strong ability to fit the training data. Specifically, Random Forest achieved 100% accuracy, precision, recall, and F1-Score on the training set. While its training performance was perfect, a significant drop in performance on the validation set strongly suggests severe overfitting. In contrast, Long Short-Term Memory (LSTM) networks, and vanilla Recurrent Neural Networks (RNNs) maintained metrics between 97% and 98% on the training set, performing well but slightly below Random Forest. Of particular note, Transformer2 (employing a class-balancing strategy) exhibited near-perfect performance on the training set, with an accuracy of 99.95%. This is likely attributable to the adopted class-balancing strategies, such as oversampling or weighted loss, which enabled the model to better learn and recognize minority classes, thereby improving overall training effectiveness.

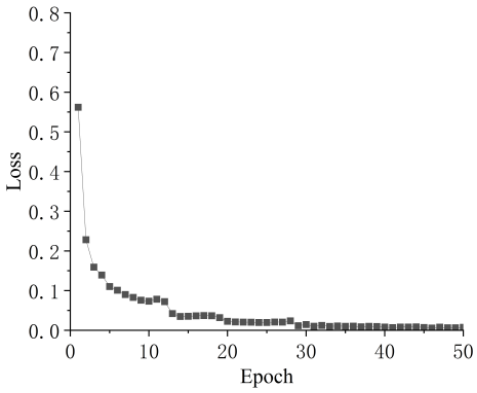
TABLE II. COMPARISON OF THE EVALUATION EFFECTS OF DIFFERENT RECOGNITION ALGORITHMS

Methods	Training process				Validation process			
	Accuracy	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Random forest	**1**	**1**	**1**	**1**	**0.9083**	**0.9095**	**0.9083**	**0.9074**
LSTM	**0.9799**	**0.9810**	**0.9799**	**0.9799**	**0.9230**	**0.9247**	**0.9230**	**0.9226**
Vanilla RNN	**0.9813**	**0.9819**	**0.9813**	**0.9813**	**0.9241**	**0.9273**	**0.9241**	**0.9232**
Transformer	**0.9776**	**0.9784**	**0.9776**	**0.9766**	**0.9247**	**0.9280**	**0.9247**	**0.9238**
Transformer2 (class-balanced)	**0.9995**	**0.9995**	**0.9995**	**0.9995**	**0.9397**	**0.9404**	**0.9397**	**0.9396**

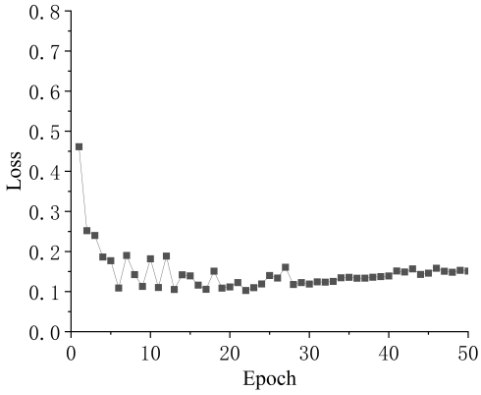
Performance on the validation set more accurately reflects a model's generalization capability—its ability to handle unseen data. Random Forest's performance significantly declined on the validation set, with an accuracy of only 90.83%, a stark contrast to its 100% training performance. This further confirms its severe overfitting, rendering the model unsuitable for direct real-world application. The LSTM model performed relatively well on the validation set, achieving an accuracy of 93.30%, outperforming vanilla RNN. This indicates that LSTMs have an advantage in processing sequential data, particularly in capturing long-term dependencies. Vanilla RNN model showed comparable performance on the validation set, with accuracies around 92.4% to 92.5%, but slightly trailed LSTM. Vanilla RNNs' performance might be limited by long-range dependency issues, while the standard Transformer's optimization might be insufficient with limited data, preventing it from fully realizing its potential in sequential data processing. Most notably, Transformer2 (class-balanced) achieved the best performance on the validation set, with an accuracy of 93.97%. Furthermore, this model also showed the smallest performance gap between the training and validation sets, dropping from 99.95% to 93.97%. This result strongly demonstrates the effectiveness of the class-balancing strategy, as it not only significantly mitigated overfitting but also substantially enhanced the model's generalization capability, allowing it to handle imbalanced datasets more effectively.

In summary, while Random Forest achieved 100% accuracy on the training set, it only managed 90.83% on the validation set, indicating severe overfitting. This might be due to the tree-based model's oversensitivity to noise in the training data. The Transformer2 (class-balanced) model demonstrated the best generalization capability, exhibiting the smallest performance gap between its training and validation sets. This highlights the crucial role of class-balancing strategies in improving a model's generalization performance on imbalanced datasets. Among other sequential models, LSTM outperformed both vanilla RNN and the standard Transformer, suggesting that on moderately sized datasets, LSTM's sequential modeling capabilities still hold a significant advantage. Furthermore, this study clearly indicates that class-balancing strategies can significantly enhance the Transformer model's generalization performance, especially in handling minority classes more favorably and robustly.

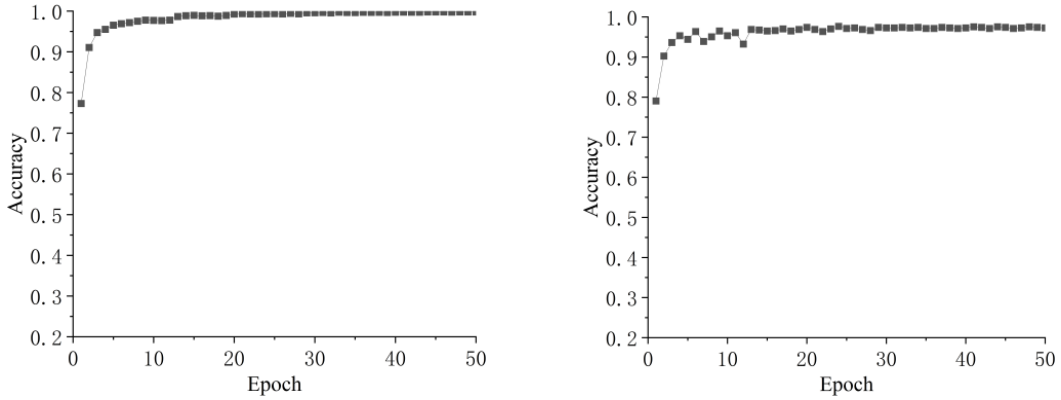
To provide a more intuitive understanding of the training and validation processes of the proposed method, relevant evaluation curves are presented in Fig. 4. As depicted in the loss curve shown in Fig. 4(a) and Fig. 4(b), the training loss value decreases rapidly with an increasing number of training epochs. By approximately 30 training epochs, the training loss value has essentially approached zero and remains highly stable thereafter. The validation loss also gradually stabilizes around 0.1 at approximately 30 training epochs. Although some fluctuations are observed subsequently, the validation loss consistently remains within the vicinity of 0.15. This indicates that the model's fit on the validation set has reached a relatively stable state.



(a) Training loss



(b) Validation loss



(c) Training accuracy (d) Validation accuracy

Fig. 4. Training and validation process of the proposed method.

Further analysis of the training accuracy curve in Fig. 4(c) reveals that the model's accuracy on the training set exceeded 0.95 when the training epochs were merely 10. It then continued to improve, gradually stabilizing and ultimately approaching 1. Similarly, as shown in the validation accuracy curve in Fig. 4(d), the model's accuracy on the validation set reached approximately 0.95 at just 10 training epochs. Despite slight subsequent fluctuations, the validation accuracy consistently remained stable around 0.95. Overall, these curves collectively demonstrate that the method proposed in this study exhibits rapid convergence in the early stages of training and achieves high performance levels on both the training and validation sets. The swift convergence of training loss to near zero, coupled with the stable fluctuations of validation loss, indicates that while the model effectively learns features from the training data, it also possesses a commendable degree of generalization capability. The consistent stability of the validation accuracy around 0.95 further corroborates the model's robustness and effectiveness.

V. Discussion and Analysis

To provide a more comprehensive analysis of the strengths and weaknesses of the proposed method, we conducted a detailed calculation of evaluation metrics for each activity class within the test set. The assessment results are shown in Table III. This approach effectively illustrates where the proposed method excels in recognition and where its performance is suboptimal. Overall, the classification model demonstrates excellent performance on the majority of classes, particularly those with higher support, but exhibits instability on categories with lower support, revealing a clear dependency on data availability.

The proposed method performs exceptionally well on high-support categories, namely Classes 1 to 6. These classes possess a relatively large number of samples, with each class having a support of over 400 samples. Across these activities, the model's precision, recall, and F1-Score are generally high. Notably, Classes 1, 2, 5, and 6 all achieved F1-Scores exceeding 90%, with Class 6 performing best at an F1-Score of 98.32%, indicating near-perfect classification. However, for Class 3, despite an exceptionally high precision of 98.88%, its recall was relatively low at 84.05%. This suggests that the model is highly strict in its predictions for this class, potentially leading to a certain degree of missed detections. To balance precision and recall, future adjustments to the classification threshold could be considered to improve recall for this category.

TABLE III. Evaluation Effect of Each Type of Human Activity

Class	Precision	Recall	F1-Score	Support degree
1	0.9215	0.9940	0.9564	496
2	0.8920	0.9654	0.9272	462
3	0.9888	0.8405	0.9086	420
4	0.9287	0.8720	0.8995	508
5	0.8798	0.9478	0.9126	556
6	1	0.9670	0.9832	545
7	0.8421	0.6957	0.7619	23
8	0.7692	1	0.8696	10
9	0.6786	0.5938	0.6333	32
10	0.7619	0.6400	0.6957	25
11	0.6875	0.6735	0.6804	49
12	0.6552	0.7037	0.6786	27

In contrast to the high-support categories, the proposed method's performance on low-support categories, namely Classes 7 to 12, was relatively poor and unstable. These classes have fewer samples, with each class having a support of less than 50 samples, leading to significant fluctuations in the model's evaluation metrics. For instance, in Class 8, although the model's recall was remarkably high at 100%, indicating its ability to capture all true positive samples, its precision was only 76.92%. This suggests a high false positive rate, where the model incorrectly identifies some samples as belonging to Class 8 when they do not. Furthermore, Class 9 exhibited low precision and recall, at 67.86% and 59.38% respectively, resulting in an F1-Score of 63.33%, the lowest among all categories. This clearly indicates that the model's discriminative ability for Class 9 is relatively weak. This instability primarily stems from insufficient data, as the model was unable to adequately learn the crucial features of these minority classes, leading to reduced generalization capabilities.

It is important to note that some categories, such as Classes 3, 8, and 12, exhibit an imbalance between precision and recall. This suggests that the model's classification thresholds may require optimization. Specifically, for categories with high precision but low recall (e.g., Class 3), the classification threshold could be appropriately lowered to improve recall while maintaining reasonable precision. Conversely, for categories with high recall but low precision (e.g., Class 8), the classification threshold could be raised to reduce false positives and enhance precision. While the model demonstrates excellent performance on data-rich categories, there remains significant room for improvement in its recognition effectiveness for small-sample categories. Future optimization efforts should focus on increasing the data volume for minority classes and further refining the model's classification strategies, for example, by adjusting classification thresholds or employing cost-sensitive learning. These measures aim to comprehensively enhance the model's overall performance, particularly its robustness in handling imbalanced datasets.

From the above analysis, it can be seen that the method proposed in this study combines the data balance strategy with the Transformer model. Compared with traditional machine learning methods, it shows significant advantages in recognition accuracy. Although Reshmi et al. [27] proposed a recognition method based on feature selection, their work did not fully consider the impact of data imbalance on model performance. Pavliuk et al. [28] focused on exploring the application of transfer learning in the task of human activity recognition and did not delve deeply into the issue of unbalanced category distribution either. In fact, the problem of data imbalance is widely present in sensor-based activity recognition tasks, which can easily lead to the model being biased towards the majority of classes and a decline in generalization ability. Therefore, it must be fully considered in the method design stage. This study effectively alleviates this problem by introducing a balance strategy, thereby enhancing the robustness and recognition accuracy of the model in real scenarios.

VI. Conclusion

This study proposes a deep learning-based smartphone physical activity recognition framework, designed to address the prevalent decline in physical activity among college students. The framework's innovation lies in its utilization of high-precision Inertial Measurement Units (IMUs) embedded in smartphones to capture multi-dimensional real-time body postures during daily activities. It then employs an advanced Transformer architecture as the core classifier. The Transformer's unique Self-Attention Mechanism effectively extracts complex spatiotemporal features from sensor data, enabling precise recognition of various physical activities (such as walking, running, and climbing stairs). This integrated approach of intelligent sensing and advanced algorithms provides a reliable and efficient technological solution for monitoring college students' physical activity.

1) After achieving 99.95% on the training set, the Transformer2 (class-balanced) model maintained an accuracy rate of 93.97% on the validation set. The minimum performance gap between the training set and the validation set strongly demonstrates its excellent generalization performance and ability to handle imbalanced datasets.

2) The recognition accuracy rate of the Transformer2 (Class Balance) model is 93.97%, which is superior to 93.30% of LSTM, approximately 92.4%-92.5% of vanilla RNN and standard Transformer, and 90.83% of Random Forest.

3) The recognition accuracy of high-support categories is very high, but low-support categories remain challenging. Our model achieves extremely high recognition accuracy for highly supported categories (categories 1-6, with over 400 samples in each category). However, for the low-support categories (7-12 categories, with less than 50 samples for each category), the performance of the model fluctuates greatly.

4) The class balance strategy effectively enhances the robustness of the Transformer model. Although this numerical improvement seems small, it indicates that the model is more robust when dealing with minority classes, significantly reduces overfitting caused by data imbalance, and improves the overall generalization performance.

5) This model has a fast convergence speed and stable verification accuracy in the early stage of training. This indicates that the model has efficient learning ability and good stability, and can achieve a high-performance level within a relatively short training time.

The proposed method remains dependent on data augmentation or supplementary samples to improve recognition performance for minority classes, without fundamentally resolving the underlying data scarcity challenge. Although the class-balancing strategy enhances model robustness when handling minority categories, the limited magnitude of overall performance improvement indicates that the current approach still exhibits insufficient efficacy in addressing severe class imbalance scenarios, suggesting significant potential for further refinement. Future research will focus on further improving the model's recognition performance for low-support activity categories. Explore more advanced data augmentation technologies to expand the sample of ethnic minorities. The meta-learning or Few-Shot learning methods are studied to enable effective learning even under sparse data. Additionally, we will consider extending this framework to more complex body activity pattern recognition and exploring its potential applications in long-term health monitoring and personalized health intervention.

Acknowledgment

The authors gratefully acknowledge the support from the 2022 Liaoning Provincial Natural Science Foundation Program Key Science and Technology Innovation Base Joint Open Fund(2022-KF-18-04), Shanghai Municipal Commission of Education (C2024090) and Scientific research project of Shanghai Vocational College of Agriculture and Forestry (KY(6)2-0000-23-13).

References

[1] S. Majumder, and M. Deen, “Smartphone sensors for health monitoring and diagnosis”, Sensors, vol. 19, pp. 216, 2019. DOI: 10.3390/s19092164

[2] H. Sarmadi, A. Entezami, K. V. Yuen, and B. Behkamal, "Review on smartphone sensing technology for structural health monitoring". Measurement, vol. 223, pp. 113716, 2023. DOI: 10.1016/j.measurement.2023.113716

[3] Y. G. Lee, W. S. Jeong, and G. Yoon, "Smartphone-based mobile health monitoring", Telemedicine and e-Health, vol. 18, pp. 585-590, 2012. DOI: 10.1089/tmj.2011.0245

[4] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions", Information Fusion, vol. 46, pp. 147-170, 2019. DOI: 10.1016/j.inffus.2018.06.002

[5] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection", IEEE Transactions on Kowledge and Data Engineering, vol. 20, pp. 1082-1090, 2018. DOI: 10.1109/TKDE.2007.190662

[6] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection". ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 61, p. 1-30, 2019. DOI: 10.1145/3328932

[7] M. Janidarmian, F. A. Roshan, K. Radecka, and Z. Zilic, "A comprehensive analysis on wearable acceleration sensors in human activity recognition", Sensors, vol. 17, pp. 529, 2017. DOI: 10.3390/s17030529

[8] S. Iloga, A. Bordat, J. Kernec, and O. Romain, "Human activity recognition based on acceleration data from smartphones using HMMs", IEEE Access, vol. 9, pp. 139336-139351, 2021. DOI: 10.1109/ACCESS.2021.3118472

[9] M. B. Del Rosario, S. J. Redmond, and N. H. Lovell, "Tracking the evolution of smartphone sensing for monitoring human movement", Sensors, vol. 15, pp. 18901-18933, 2015. DOI: 10.3390/s150818901

[10] I. A. Faisal, T. W. Purboyo, and A. S. R. Ansori, "A review of accelerometer sensor and gyroscope sensor in IMU sensors on motion capture". ARPN Journal of Engineering and Applied Science, vol. 15, pp. 826-829, 2019. DOI: 10.36478/jeasci.2020.826.829

[11] I. M. Pires, N. M. Garcia, E. Zdravevski, and P. Lameski, "Daily motionless activities: a dataset with accelerometer, magnetometer, gyroscope, environment, and GPS data", Scientific Data, vol. 9, pp. 105, 2022. DOI: 10.1038/s41597-022-01213-9

[12] S. Hernandez Sanchez, R. Fernandez Pozo, and L. A. Hernandez Gomez, "Estimating vehicle movement direction from smartphone accelerometers using deep neural networks", Sensors, vol. 18, pp. 2624, 2018. DOI: 10.3390/s18082624

[13] X. Huang, Y. Xue, S. Ren, and F. Wang, "Sensor-based wearable systems for monitoring human motion and posture: A review", Sensors, vol. 23, pp. 9047, 2023. DOI: 10.3390/s23229047

[14] O. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data. Pervasive and Mobile Computing", vol. 8, pp. 717-729, 2012. DOI: 10.1016/j.pmcj.2012.06.004

[15] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones", Applied Sciences, vol. 7, pp. 1101, 2017. DOI: 10.3390/app7101101

[16] M. O. Mario, "Human activity recognition based on single sensor square HV acceleration images and convolutional neural networks". IEEE Sensors Journal, vol. 19, pp. 1487-1498, 2018. DOI: 10.1109/JSEN.2018.2882943

[17] W. Niu, J. Long, D. Han, and Y. Wang, "Human activity detection and recognition for video surveillance", In 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), Taipei, pp. 719-722, 2014. DOI: 10.1109/ICME.2004.1394293

[18] J. Sung, C, Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images", In 2012 IEEE International Conference on Robotics and Automation, Saint Paul, p. 842-849, 2012. DOI: 10.1109/ICRA.2012.6224591

[19] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection", IEEE Transactions on Cybernetics, vol. 43, pp. 1383-1394, 2013. DOI: 10.1109/TCYB.2013.2276433

[20] H. Koppula, and A. Saxena, "Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation", In Proceedings of the 30th International Conference on Machine Learning, Atlanta, pp. 792-800, 2013. https://proceedings.mlr.press/v28/koppula13.html

[21] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic petri net framework for human activity detection in video", IEEE Transactions on Multimedia, vol. 10, pp. 1429-1443, 2008. DOI: 10.1109/TMM.2008.2010417

[22] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection", In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, pp. 2626-2634, 2022. DOI: 10.1109/WACV45572.2020.9093633

[23] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models", Journal on Artificial Intelligence, vol. 6, pp. 301-360, 2024. DOI: 10.32604/jai.2024.054314

[24] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning", Neurocomputing, vol. 452, pp. 48-62, 2021. DOI: 10.1016/j.neucom.2021.03.091

[25] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer", Advances in Neural Information Processing Systems, vol. 34, pp. 15908-15919, 2021. https://arxiv.org/abs/2103.00112.

[26] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones", In ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Belgium, pp. 437-442, 2013.

[27] S. Reshmi, & E. Ramanujam, "An ensemble maximal feature subset selection for smartphone based human activity recognition", Journal of Network and Computer Applications, vol. 226, pp. 103875, 2024. DOI: 10.1016/j.jnca.2024.103875

[28] O. Pavliuk, M. Mishchuk, and C. Strauss, "Transfer learning approach for human activity recognition based on continuous wavelet transform", Algorithms, vol. 16, no.2, pp. 77, 2023. DOI: 10.3390/a16020077