A Key Feature Screening Method for Human Activity Recognition Based on Multi-head Attention Mechanism

Hao Wang[1*] Fangyu Liu[1,2*] Xiang Li[3] Ye Li[1] Fangmin Sun[1†]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]School of Electronics and Information Engineering, South China Normal University, Foshan, China
{h.wang, fy.liu1, xiang.li2, ye.li, fm.sun}@siat.ac,cn

Abstract

Human activity recognition using wearable sensors has become a critical task in ubiquitous computing, with applications in healthcare, fitness monitoring, and smart environments. However, sensor-based HAR often involves high-dimensional, multi-channel time series data, where redundant or irrelevant features may degrade both classification performance and model interpretability. In this study, we propose a lightweight feature screening framework guided by a multi-head attention mechanism to address this challenge. The model first applies channel-wise linear transformations to extract localized representations from each sensor axis, and then employs a multi-head attention module to dynamically assess the importance of each feature across all channels. This design enables the model to emphasize the most informative components while suppressing noise and redundancy. Experiments on the KU-HAR dataset demonstrate that the proposed method achieves 96.0% classification accuracy using only 60 selected features. In addition, the selected features provide valuable references for future research in feature selection, model simplification and multimodal sensor fusion.

1. Introduction

Human Activity Recognition (HAR) refers to the process of automatically identifying and classifying human actions based on data collected from various sensors [1]. HAR has been widely applied in various fields, including sports training, healthcare, and security monitoring [2]. In sports training, HAR technology is used to monitor athlete movements and optimize performance while minimizing the risk of injury [3]. In healthcare, HAR supports patient health monitoring and rehabilitation assessment by tracking movement patterns, which helps in early detection of conditions such as weakness or falls [4]. In the field of security, HAR helps to automatically detect abnormal behaviors and improve public safety in sensitive areas [5]. With the rapid development of artificial intelligence and the Internet of Things (IoT), HAR has become an increasingly important research field.

Human activities are complex and variable, exhibiting diverse temporal dynamics and motion patterns. This fine-grained distinction makes activity recognition a challenging task. To address this challenge, researchers have explored multiple sensing modalities for HAR, primarily including vision-based methods [6], wireless signal-based methods [7], and wearable sensor-based methods [8]. Each approach offers unique advantages but faces limitations due to environmental sensitivity, data quality, or practical deployment constraints.

Vision-based HAR methods use cameras to capture the spatial and temporal features of human motion and then analyze human poses and actions. These approaches often achieve high accuracy in controlled environments by leveraging advanced computer vision models. However, vision-based methods are susceptible to environmental conditions such as illumination changes, occlusions, and camera placement. Moreover, video surveillance raises many privacy concerns that limit their application in healthcare or daily life scenarios [9].

Wireless signal-based HAR methods exploit variations in wireless signals, such as Wi-Fi channel state information (CSI) or radio frequency (RF) reflections, to recognize human activities. These methods can classify activities without relying on visual or wearable data. However, the recognition accuracy of wireless-based HAR is highly dependent on environmental layout, including furniture layout, walls, and signal interference, and fine-grained activity discrimination remains challenging due to the coarse nature of signal reflections [10].

[*]Equal contribution
[†]Corresponding author

In contrast, wearable sensor-based HAR methods have attracted attention due to their ability to capture physiological and motion-related signals directly from the human body. A wide range of wearable sensors can be used for activity monitoring, including inertial measurement unit (IMU), electromyography (EMG) sensors, electrocardiography (ECG) sensors, pressure sensors, and skin temperature sensors. Among them, IMU typically consists of a three-axis accelerometer and a gyroscope, which are used to measure linear and angular motion in multiple directions. It is usually embedded in smartphones, smartwatches or dedicated wearable devices, and can achieve continuous and inconspicuous activity recognition in various environments.

Despite their advantages, IMU-based HAR systems still face several challenges. First, the high-dimensional and multi-channel nature of IMU data increases computational burden and model complexity. Second, not all extracted features contribute equally to recognition performance; irrelevant or redundant features can degrade classification accuracy. Third, most deep learning or end-to-end models used in HAR provide limited interpretability, offering little insight into which features or sensor channels contribute most to the final prediction.

To address these challenges, we propose a key feature screening method based on a multi-head attention mechanism.The method first employs independent linear transformations for each sensor channel to capture the unique information embedded in handcrafted features from each channel. A multi-head self-attention module is then introduced to evaluate the importance of each feature channel adaptively. The model generates interpretable global importance scores for effective feature selection by averaging attention weights across all heads and validation samples. Experiments on KU-HAR datasets demonstrate that the proposed method achieves competitive classification accuracy using only a subset of the most informative features, highlighting both its efficiency and interpretability. Unlike traditional feature selection methods that require computing correlation coefficients or mutual information for each parameter, our method leverages multi-head attention to simultaneously identify attention-worthy features from a large feature set in an efficient and scalable manner. The main contributions are summarized as follows:

1. We propose a novel feature screening framework for wearable-sensor-based HAR, utilizing multi-head attention to compute global importance scores and rank features across all channels.

2. Our method combines independent channel-wise linear transformations with attention-guided feature selection, producing a compact and highly informative feature set that enhances both classification performance and interpretability.

3. Experimental results demonstrate that our attention-based feature selection consistently outperforms full-feature deep learning baselines and traditional filter-based selection methods, highlighting its efficiency, scalability, and practical value for resource-constrained deployment.

The remainder of this paper is organized as follows. Section 2 presents related work, including traditional handcrafted feature extraction and deep learning-based models for HAR. Section 3 details the dataset, feature extraction, and the proposed model architecture. Section 4 describes the experimental setup and performance evaluation on public datasets. Finally, Section 5 concludes the study and outlines directions for future work.

2. Related Work

In this section, we review the most relevant literature from two perspectives: traditional handcrafted feature extraction, deep learning-based HAR models applied in sensor data modeling.

2.1. Traditional Feature Extraction in Human Activity Recognition

Traditional human activity recognition methods primarily relied on handcrafted feature extraction from time series signals, such as statistical, temporal, and frequency domain features [11]. These features are usually based on expertise, extracted by sliding window techniques, and fed into traditional machine learning algorithms for classification, such as support vector machine (SVM) [12], random forest (RF) [13], k-nearest neighbors (KNN) [14] and decision tree (DT) [15]. With the development of time series feature extraction tools, tools like TSFEL [11] and TSFresh [16] have been developed to automate the process, enabling the extraction of hundreds of features covering a wide range of signal properties.

Despite their effectiveness in simple tasks, handcrafted features may not to capture complex, non-linear relationships in human activity. In addition, the large number of handcrafted features can introduce redundancy and noise, reducing the model's generalizability.

2.2. Deep Learning for Human Activity Recognition

With the development of deep learning, traditional machine learning methods have been replaced by deep learning methods. Deep learning offers a powerful, data-driven alternative to handcrafted feature extraction by enabling end-to-end modeling of complex time series signals. Convolutional Neural Network (CNN) are widely used to extract local temporal patterns from sensor data [17, 18], while Recurrent Neural Networks (RNN), including LSTM and

GRU, are capable of modeling long-term temporal dependencies [19, 20].

To further enhance performance, attention mechanisms have been integrated into HAR models to dynamically focus on informative segments of input data. Temporal attention enables the model to prioritize important time steps, hile channel- or modality-level attention helps leverage complementary information across multiple sensor streams. For example, Mekruksavanich et al. [21] proposed a hybrid CNN incorporating channel attention mechanism, called ResNet-BiGRU-SE, to extract deep spatio-temporal features. Stuchbury-Wass et al. [22] also demonstrated that attention mechanisms can improve model robustness and interpretability in wearable HAR applications.

However, many attention-based models still lack explicit feature selection and offer limited interpretability at the feature or channel level. Attention is often used internally to improve learning, but rarely for feature screening or reducing input dimensionality. In this study, we propose a multi-head attention-based framework that directly evaluates feature importance across channels, enabling compact, interpretable, and accurate activity recognition.

3. Materials and Methods

3.1. Dataset Description

In this study, we use the KU-HAR dataset [23], which is a public dataset for human activity recognition collected from 90 subjects (75 male and 15 female) aged between 18 and 34 years. The activity data were collected using the built-in accelerometer and gyroscope sensors of the smartphone worn at the waist. The dataset consists of 1945 raw activity samples representing 18 different classes, including simple postures (such as standing, sitting) and complex movements (such as stair climbing, playing table tennis). Each raw sample contains six tri-axial data channels: three from the accelerometer and three from the gyroscope.

3.2. Feature Extraction

The KU-HAR dataset provides 20750 sub-samples extracted from the original samples, each sample was segmented into 3-second non-overlapping windows to ensure temporal consistency. Each window contains six sensor channels: three from the accelerometer and three from the gyroscope. We used the Time Series Feature Extraction Library (TSFEL), which is a Python package designed to extract features efficiently and comprehensively from time series data [11].

TSFEL offers a wide range of feature extraction methods in various fields. In this work, we use TSFEL to extract a total of 156 handcrafted features from each of the six sensor channels, including 31 statistical features (such as mean, variance, and skewness), 14 temporal features (such as zero-crossing rate and slope), and 111 spectral features (such as spectral centroid and bandwidth). This results in a total of 936 features per sample, capturing comprehensive time, frequency, and statistical information from the raw IMU signals. The formulas for 156 features can be found in Paper [11].

3.3. Model Architecture

The goal of our approach is to achieve both high recognition accuracy and model interpretability for wearable sensor-based HAR. Unlike conventional end-to-end models that typically utilize all extracted features without explicit feature selection, our method introduces an explicit feature screening stage, powered by a multi-head attention mechanism. This design enables our model to directly quantify and rank the global importance of each feature parameter, so that only the most informative features are selected for downstream classification. Such explicit attention-guided feature selection not only improves computational efficiency but also provides a clear interpretation of which features contribute most to the HAR task. The proposed framework [1] is depicted in Figure. 1 and the detailed computational flow is summarized in Algorithm 1.

Algorithm 1 A key feature screening method for HAR.

Input: The waist IMU signal set $\mathcal{S} = \{s_n\}_{n=1}^{N}$ of all samples.

Output: Human activity recognition result y and the feature wise attention weights α of the multi-head attention module.

```
for $s \in \mathcal{S}$ do
    Use the TSFEL library to extract the signal's statistical, temporal, and spectral features $x \in X$.
end for
$Z_0 = X \in \mathbb{R}^{B \times C \times D_{dim}}$
$Z_l = Einsum(Z_{l-1}, W_l) + b_l$                //Linear Layer
$Q = Z_l * W_Q + b_Q$                             //Query
$K = Z_l * W_K + b_K$                             //Key
$V = Z_l * W_V + b_V$                             //Value
$Scores = \frac{QK^T}{\sqrt{d_k}} \in \mathbb{R}^{B \times H \times C \times C}$    //H:Head
$Weights : \alpha = Softmax(Scores, dim = -1)$
$Out_{Attn} = \alpha * V * W_{out} + b_{out}$
$y = Softmax(Dropout(Out_{Attn}) * W_{cls} + b_{cls})$
return $y, \alpha$                  //$\alpha$ guide to feature screening
```

1) Independent Linear Transformations:

Given an input sample $X \in \mathbb{R}^{B \times C \times D_{dim}}$, where B is the batch size, C is the number of feature channels, and D_{dim} is the feature dimension per channel. To preserve the independent semantics of each feature channel, we use two independent linear layers for per-channel transformation.

[1] Code link: https://github.com/FangyuLiu-2023/AKFSM

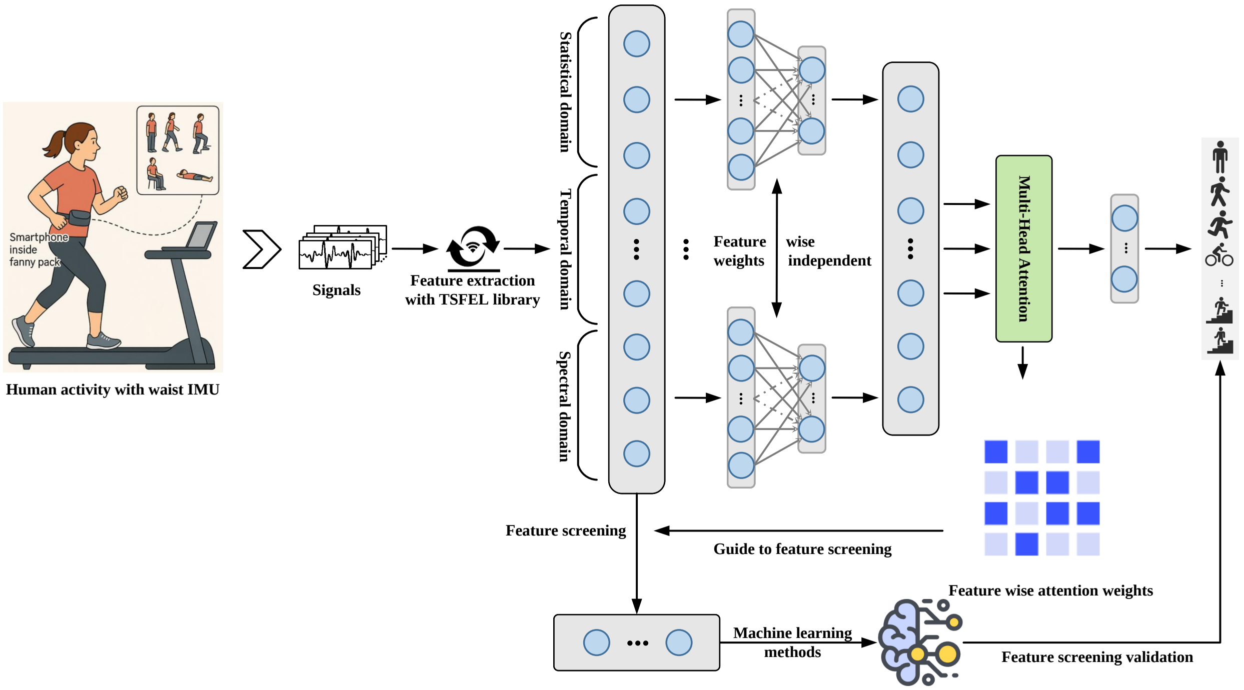

Smartphone inside fanny pack
Human activity with waist IMU
Signals
Feature extraction with TSFEL library
Statistical domain
Temporal domain
Spectral domain
Feature weights
wise independent
Multi-Head Attention
Feature screening
Guide to feature screening
Feature wise attention weights
Machine learning methods
Feature screening validation

Figure 1. Key feature screening network architecture for HAR based on multi-head attention mechanism.

These layers are computed using operations, with learnable weight tensors, W_1 , W_2 and biases. This design enables each feature channel to be transformed independently, which reduces computational overhead while maintaining channel independence.

2) Multi-head Attention for Feature Importance Scoring and Selection:

The transformed features are passed to a multi-head self-attention module, which captures inter-channel dependencies by computing query, key, and value vectors for each head. The attention scores are calculated using scaled dot-product, and softmax is applied to obtain attention weights. These weights are used to re-weight the value vectors, allowing the network to selectively focus on the most informative features. This module enables the model to focus on key feature channels and enhances the feature interactions that are critical for activity classification.

3) Final Classification Layer:

The final attended features are flattened and passed through a dropout layer and a linear classifier to generate the predicted activity label $y \in \mathbb{R}^{B \times N}$, where $N = 18$ is the number of activity classes.

4. Experiment and Analysis

4.1. Experimental Setup

To evaluate the performance of the proposed feature selection method based on multi-head attention, we conduct experiments on the KU-HAR dataset. Following the preprocessing procedure in Section 3.2, we used the TSFEL library to extract a rich set of time series features, which are then used as input to the attention-based model and other comparison baselines.

To ensure a comprehensive and interpretable evaluation of the selected features, we employ a diverse set of traditional classifiers, including SVM, RF, DT, KNN, Logistic Regression (LR), Gradient Boosting (GB), and Light Gradient Boosting Machine (LightGBM). These classifiers represent different types of classification paradigms, such as distance-based (KNN), linear models (LR, SVM), tree-based methods (DT, RF), and boosting algorithms (GB, LightGBM).

We adopted a five-fold cross-validation strategy to test the proposed method and compared it with several representative HAR models, including:

DenseNet-GRU [25]: a hybrid architecture combining dense convolutional blocks and gated recurrent units.

CNN [18]: a standard convolutional neural network for local feature learning.

ResRNN [26]: a residual recurrent model designed for capturing sequential dependencies.

CNN-LSTM [27]: a widely used architecture integrating convolutional feature extractors and temporal modeling via LSTM.

Multi-STMT [28]: a multilevel network based on the spatiotemporal attention mechanism and multiscale temporal embedding.

Table 1. The top ten important features were selected.

Feature name	Domain	Description	Formula
MFCC_9	Spectral	Mel-scale frequency cepstral coefficients	As in paper [24]
Spectral distance	Spectral	The signal spectral distance	$\sum_{i=0}^{N} lr_{fmag_i} - cumsum_{fmag_i}$
Positive turning points	Temporal	Number of positive turning points of the signal	$\sum_{i=0}^{N-2} \mathbf{1}(\frac{ds_i}{dt} > 0 \wedge \frac{ds_{i+1}}{dt} < 0)$
Maximum frequency	Spectral	Maximum frequency of the signal	$freq[min\{i \| cumsum_{fmag_i} \geq$ $0.95 \cdot cumsum_{fmag_{max}}\}]$
ECDF Percentile Count_1	Statistical	The cumulative sum of samples that are less than the percentile	$\sum_{i=0}^{N} \mathbf{1}(ECDF\ values(s_i) < p)$
Signal distance	Temporal	Signal traveled distance	$\sum_{i=0}^{N-1} \sqrt{1 + \Delta s_i^2}$
Spectral positive turning points	Spectral	The number of positive turning points of the fft magnitude signal	$\sum_{i=0}^{N-2} \mathbf{1}(\frac{dfmag_i}{dfreq_i} > 0 \wedge \frac{dfmag_{i+1}}{dfreq_{i+1}} < 0)$
Negative turning points	Temporal	Number of negative turning points of the signal	$\sum_{i=0}^{N-2} \mathbf{1}(\frac{ds_i}{dt} < 0 \wedge \frac{ds_{i+1}}{dt} > 0)$
Power bandwidth	Spectral	Power spectrum density bandwidth of the signal	$\|max\{freq \| C(freq) \leq 0.95 \cdot C(freq_{max})\} -$ $min\{freq \| C(freq) \geq 0.95 \cdot C(freq_{max})\}\|$
Zero crossing rate	Temporal	Zero-crossing rate of the signal	$\sum_{i=0}^{N-1} \mathbf{1}(sign(s_i) \neq sign(s_{i+1}))$

s: signal vector, Δs: the discrete derivative of s, t: time vector, fs: signal's sampling frequency, N: the length of s, $freq, fmag = fft(t, s)$, $cumsum$: cumulative sum, lr: linear regression, $\mathbf{1}\{\cdot\}$: indicator function, p: percentile value, $C(freq) = \sum_{i=0}^{freq} |fmag_i|^2$, $ECDF$: empirical cumulative distribution function.

ResNet-BiGRU-SE [21]: a hybrid CNN model embedded with a channel attention mechanism (squeeze-and-excitation block) to hierarchically extract deep spatio-temporal features.

All models are implemented in PyTorch and trained using the Adam optimizer. The initial learning rate is set to 10^{-4}. We evaluate the model performance using six commonly used metrics for multi-class classification, including accuracy (ACC), precision (PRE), recall (REC), F1-score (F1), Matthews correlation coefficient (MCC), and area under the ROC curve (AUC).

4.2. Feature Screening and Analysis

Although existing HAR methods often rely on using all extracted features without discrimination, we believe that not all features contribute equally to the classification task. To improve the interpretability and efficiency of the model, we analyze the contribution of individual features by utilizing the attention weights produced by the multi-head attention module.

We compute the global importance scores to find the most informative features by averaging the attention weights across all heads and validation samples. This method generates an importance value for each feature channel c, which is calculated as follows:

$$\bar{\alpha} = \frac{1}{B \cdot H} \sum_{b=1}^{B} \sum_{h=1}^{H} \alpha_{b,h,c}$$

where B is the number of samples in the validation set, H is the number of attention heads, and $\alpha_{b,h,c}$ is the attention weight of channel c in head h for sample b. Based on these scores, we rank the features and select the top-k most informative features. In our implementation, k is set to 10, and since there are 6 sensor channels, this results in a total of 60 selected features. Table 1 lists the top 10 selected features.

4.3. Classification Performance

To evaluate the effectiveness of the proposed attention-based model on the HAR task, we compare its performance with several deep learning methods. As shown in Table 2, the proposed model outperforms most existing methods across all metrics, achieving an average accuracy, precision, recall, and F1-score of 93%, with a relatively low number of parameters (0.79M) and FLOPs (1.17M). Compared with larger models such as DenseNet-GRU and Multi-STMT, The feature screening model we proposed is significantly lighter weight while maintaining high accuracy, demonstrating its suitability for deployment on resource-constrained wearable devices.

In addition, we use the features selected by the attention mechanism are used to test a variety of traditional machine learning classifiers, including LR, DT, KNN, RF, SVM, GB and LightGBM. Figure 2 and Table 2 shows that LightGBM achieves the best results across all metrics, with an accuracy and F1-score of 96.0%, and an AUC of 0.999, outperforming several deep learning baselines. Other models such as Random Forest, Gradient Boosting, and Support Vector Machine also achieve accuracies above 90%. The results demonstrate that we achieve higher performance by using a small number of selected features and machine learning methods. The potential reason for this improvement is that the attention mechanism emphasizes informative and task-relevant features, while automatically filtering out redundant or noisy features. Therefore, the filtered features still have strong discriminative power, enabling accurate and efficient classification.

Table 2. Results of Experiment.

Year	Method	Data	Human activity recognition							
			KU-HAR Dataset: 20,750 samples from 90 subjects (75 Male / 15 Female)							
			ACC	PRE	REC	F1	MCC	AUC	FLOPs	Params
2021	DenseNet-GRU [25]	Waist IMU	0.89±0.01	0.89±0.01	0.89±0.01	0.89±0.01	0.88±0.01	0.97±0.00	54.53M	1.31M
2022	CNN [18]	Waist IMU	0.83±0.02	0.84±0.01	0.83±0.02	0.82±0.02	0.82±0.02	0.98±0.00	3.28M	1.19M
2022	ResRNN [26]	Waist IMU	0.76±0.01	0.76±0.06	0.76±0.01	0.71±0.02	0.76±0.01	0.90±0.02	17.19M	1.29M
2023	ResNet-BiGRU-SE [21]	Waist IMU	0.89±0.01	0.90±0.01	0.89±0.01	0.89±0.01	0.89±0.01	0.99±0.00	0.08G	4.06M
2024	CNN-LSTM [27]	Waist IMU	0.80±0.01	0.82±0.02	0.80±0.01	0.80±0.01	0.79±0.01	0.97±0.00	7.05M	1.85M
2024	Multi-STMT [28]	Waist IMU	0.85±0.01	0.87±0.02	0.85±0.01	0.85±0.01	0.84±0.01	0.98±0.01	47.70M	5.35M
Ours	Linear+Attention	All Features	**0.93±0.01**	**0.93±0.01**	**0.93±0.01**	**0.93±0.01**	**0.93±0.01**	0.90±0.02	**1.17M**	**0.79M**
Ours	LR	Select Features	0.81±0.00	0.81±0.00	0.81±0.00	0.81±0.00	0.80±0.00	0.99±0.00	1.05K	**1.07K**
Ours	DT	Select Features	0.83±0.00	0.83±0.00	0.83±0.00	0.83±0.00	0.82±0.00	0.90±0.00	3.00K	5.99K
Ours	KNN	Select Features	0.78±0.00	0.78±0.00	0.78±0.00	0.77±0.01	0.76±0.01	0.96±0.00	**0.59K**	0.95M
Ours	RF	Select Features	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.92±0.00	1.00±0.00	0.33M	0.66M
Ours	SVM	Select Features	0.85±0.01	0.85±0.01	0.85±0.01	0.85±0.01	0.84±0.01	0.99±0.00	0.56M	0.56M
Ours	GB	Select Features	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.92±0.00	1.00±0.00	1.45K	2.90K
Ours	LightGBM	Select Features	**0.96±0.00**	**0.96±0.00**	**0.96±0.00**	**0.96±0.00**	**0.95±0.00**	**1.00±0.00**	0.26M	0.51M

Tablenotes: FLOPs is the computational cost per sample inference

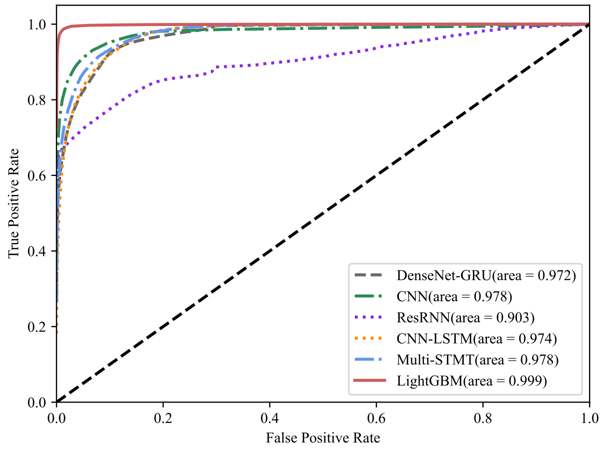


Figure 2. Illustration of the ROC curves obtained by different methods.

The class-wise classification results are visualized in Figure 3 (b). We observe that activities such as Jump, Pick, and Lay-stand achieve consistently high F1-scores, indicating that the proposed model can effectively distinguish these actions with significant differences. In contrast, activities with more ambiguous or overlapping motion patterns, such as Walk-circle and Walk-backward, show relatively lower performance. This might be because the movement patterns among these activities are very similar.

As shown in Figure 3 (a), the global average confusion matrix demonstrates that most predictions are concentrated along the diagonal, indicating high overall classification accuracy. The confusion between Talk-Stand, sit and lying is particularly notable, which may be attributed to their similar low-movement signal profiles.

These results demonstrate that the proposed model not only achieves competitive classification performance, but also exhibits better computational efficiency and generalization compared to several state-of-the-art HAR models.

4.4. The Impact of Model Components

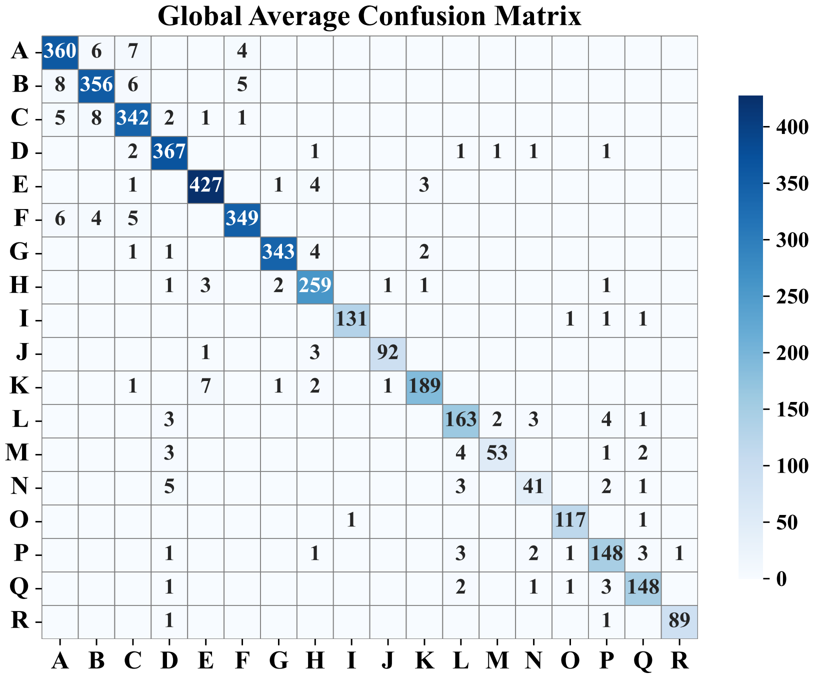
To further investigate the contribution of different signal components to the overall recognition performance, we perform an ablation study by systematically removing individual sensor axes from the input feature set. We evaluate the model's performance when excluding one of the six channels, as well as when removing entire sensors (all accelerometer or all gyroscope channels).

The results are presented in Table 3, which shows the classification performance under each ablation condition using the LightGBM classifier trained on the selected features. It can be observed that the removal of individual accelerometer axes has only a marginal effect on performance, with accuracy remaining above 95% in all cases. In contrast, removing any single gyroscope axis leads to a more noticeable drop, particularly when the x-axis of the gyroscope is excluded. When the entire gyroscope signal is removed, the classification accuracy drops significantly to 87%, while removing the entire accelerometer signal results in a relatively smaller decline to 92%. These results suggest that, within the waist-mounted IMU setup, gyroscope data plays a more critical role than accelerometer data for distinguishing human activities.

It highlights the dominant role of gyroscope features in HAR tasks using the KU-HAR dataset and supports the effectiveness of feature-level interpretability enabled by our attention-based screening method.

5. Conclusion

In this paper, we proposed a lightweight and interpretable feature screening framework for HAR using wear-



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
A	360	6	7			4												
B	8	356	6			5												
C	5	8	342	2	1	1												
D			2	367				1				1	1	1		1		
E			1		427		1	4			3							
F	6	4	5			349												
G			1	1			343	4			2							
H				1	3		2	259		1	1					1		
I									131						1	1	1	
J					1			3		92								
K			1		7		1	2		1	189							
L				3								163	2	3		4	1	
M				3								4	53			1	2	
N				5								3		41		2	1	
O									1						117		1	
P				1				1				3		2	1	148	3	1
Q				1								2		1	1	3	148	
R				1												1		89

Class	PRE	REC	F1
Stand	0.952	0.954	0.953
Sit	0.955	0.950	0.952
Talk-sit	0.937	0.952	0.944
Talk-stand	0.955	0.982	0.969
Stand-sit	0.973	0.981	0.977
Lay	0.970	0.962	0.966
Lay-stand	0.986	0.973	0.979
Pick	0.947	0.971	0.959
Jump	0.991	0.980	0.986
Push-up	0.977	0.958	0.967
Sit-up	0.973	0.940	0.956
Walk	0.924	0.926	0.925
Walk-backward	0.947	0.842	0.891
Walk-circle	0.862	0.792	0.824
Run	0.973	0.982	0.977
Stair-up	0.921	0.929	0.924
Stair-down	0.940	0.945	0.943
Table-tennis	0.980	0.972	0.976
Avg=‘Weighted’	0.958	0.957	0.957

Figure 3. (a) The five-fold global average confusion matrix and (b) class-wise performace of the classification operation.

Table 3. Impact of components in LightGBM on the classification performance of human activity recognition.

Components	Human activity recognition							
	KU-HAR Dataset: 20,750 samples from 90 subjects (75 Male / 15 Female)							
	ACC	PRE	REC	F1	MCC	AUC	FLOPs	Params
Without Accelerometer x axes	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	1.00±0.00	0.28M	0.56M
Without Accelerometer y axes	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	1.00±0.00	0.27M	0.54M
Without Accelerometer z axes	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	0.94±0.00	1.00±0.00	0.29M	0.58M
Without Gyro x axes	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	1.00±0.00	0.31M	0.63M
Without Gyro y axes	0.94±0.00	0.94±0.00	0.94±0.00	0.94±0.00	0.94±0.00	1.00±0.00	0.29M	0.59M
Without Gyro z axes	0.95±0.00	0.95±0.00	0.95±0.00	0.95±0.00	0.94±0.00	1.00±0.00	0.28M	0.56M
Without Accelerometer	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.91±0.00	1.00±0.00	0.38M	0.76M
Without Gyro	0.87±0.01	0.87±0.01	0.87±0.01	0.87±0.01	0.86±0.01	0.99±0.00	0.46M	0.92M
ALL	**0.96±0.00**	**0.96±0.00**	**0.96±0.00**	**0.96±0.00**	**0.95±0.00**	**1.00±0.00**	**0.26M**	**0.51M**

Tablenotes: FLOPs is the computational cost per sample inference

able sensor data. By combining channel-wise linear transformation with a multi-head attention mechanism, the proposed model is capable of dynamically evaluating the importance of each feature and focusing on the most relevant components while suppressing redundant information. Experiments conducted on the KU-HAR dataset show that, compared with the state-of-the-art deep learning models, our method achieves a higher classification accuracy. The LightGBM classifier trained on the selected 60 features achieved a classification accuracy rate of 96.0%, which is superior to using all features or deep learning models. The selected features serve as valuable references for future studies on feature selection strategies, lightweight model design, and multimodal sensor fusion.

In the future, we plan to extend this work by exploring adaptive feature selection strategies across different sensor placements and subject profiles, as well as investigating temporal attention mechanisms for modeling longer-term dependencies in continuous activity sequences.

6. Acknowledgments

This work was supported in part by the Key Research and Development Plan of Guangdong Province under Grant 2022B1515120062; in part by the Shenzhen International Cooperation Project under Grant GJHZ20220913142808016; in part by the Shenzhen Sustainable Development Special Project under Grant KCXFZ20230731094100001 and

KCXFZ20240903094300001; in part by the Joint Fund of NSFC and Chongqing under Grant U21A20447; and in part by the Joint Fund of Yeqisun and NSFC under Grant U2241210; in part by Guangzhou Key R&D Program under Grant 202206010127 and 2023B03J1341, and in part by National Natural Science Foundation of China under Grant 62401215.

References

[1] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: a narrative review," *Artificial intelligence review*, vol. 55, no. 6, pp. 4755–4808, 2022.

[2] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–34, 2021.

[3] O. Barut, L. Zhou, and Y. Luo, "Multitask lstm model for human activity recognition and intensity estimation using wearable sensor data," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8760–8768, 2020.

[4] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things," *Information Fusion*, vol. 94, pp. 17–31, 2023.

[5] A. Hussain, S. U. Khan, N. Khan, M. Shabaz, and S. W. Baik, "Ai-driven behavior biometrics framework for robust human activity recognition in surveillance systems," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107218, 2024.

[6] Y. Ma and R. Wang, "Relative-position embedding based spatially and temporally decoupled transformer for action recognition," *Pattern Recognition*, vol. 145, p. 109905, 2024.

[7] X. Lu, L. Wang, C. Lin, X. Fan, B. Han, X. Han, and Z. Qin, "Autodlar: a semi-supervised cross-modal contact-free human activity recognition system," *ACM Transactions on Sensor Networks*, vol. 20, no. 4, pp. 1–20, 2024.

[8] M.-K. Yi, W.-K. Lee, and S. O. Hwang, "A human activity recognition method based on lightweight feature extraction combined with pruned and quantized cnn for wearable device," *IEEE Transactions on Consumer Electronics*, vol. 69, no. 3, pp. 657–670, 2023.

[9] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[10] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3200–3225, 2022.

[11] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "Tsfel: Time series feature extraction library," *SoftwareX*, vol. 11, p. 100456, 2020.

[12] Y. Tian, X. Wang, W. Chen, Z. Liu, and L. Li, "Adaptive multiple classifiers fusion for inertial sensor based human activity recognition," *Cluster Computing*, vol. 22, pp. 8141–8154, 2019.

[13] C. Dewi and R.-C. Chen, "Human activity recognition based on evolution of features selection and random forest," in *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 2019, pp. 2496–2501.

[14] S. Mohsen, A. Elkaseer, and S. G. Scholz, "Human activity recognition using k-nearest neighbor machine learning algorithm," in *Proceedings of the International Conference on Sustainable Design and Manufacturing*. Springer, 2021, pp. 304–313.

[15] G. De Leonardis, S. Rosati, G. Balestra, V. Agostini, E. Panero, L. Gastaldi, and M. Knaflitz, "Human activity recognition by wearable sensors: Comparison of different classifiers for real-time applications," in *2018 IEEE international symposium on medical measurements and applications (memea)*. IEEE, 2018, pp. 1–6.

[16] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[17] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, "Metasense: few-shot adaptation to untrained conditions in deep mobile sensing," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 110–123.

[18] A. Dahou, M. A. Al-qaness, M. Abd Elaziz, and A. Helmi, "Human activity recognition in ioht applications using arithmetic optimization algorithm and deep learning," *Measurement*, vol. 199, p. 111445, 2022.

[19] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, “A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare,” *Information Fusion*, vol. 55, pp. 105–115, 2020.

[20] L. Wang and R. Liu, “Human activity recognition based on wearable sensor using hierarchical deep lstm networks,” *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 837–856, 2020.

[21] S. Mekruksavanich and A. Jitpattanakul, “Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition,” *Scientific Reports*, vol. 13, no. 1, p. 12067, 2023.

[22] J. Stuchbury-Wass, A. Ferlini, and C. Mascolo, “Multimodal attention networks for human activity recognition from earable devices,” in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 258–260.

[23] N. Sikder and A.-A. Nahid, “Ku-har: An open dataset for heterogeneous human activity recognition,” *Pattern Recognition Letters*, vol. 146, pp. 46–54, 2021.

[24] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[25] H. Eom, J. Roh, Y. S. Hariyani, S. Baek, S. Lee, S. Kim, and C. Park, “Deep learning-based optimal smart shoes sensor selection for energy expenditure and heart rate estimation,” *Sensors*, vol. 21, no. 21, p. 7058, 2021.

[26] M. A. Al-Qaness, A. M. Helmi, A. Dahou, and M. A. Elaziz, “The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis,” *Biosensors*, vol. 12, no. 10, p. 821, 2022.

[27] V. M. P. Cortes, A. Chatterjee, and D. Khovalyg, “Dynamic personalized human body energy expenditure: Prediction using time series forecasting lstm models,” *Biomedical Signal Processing and Control*, vol. 87, p. 105381, 2024.

[28] H. Zhang and L. Xu, “Multi-stmt: Multi-level network for human activity recognition based on wearable sensors,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.