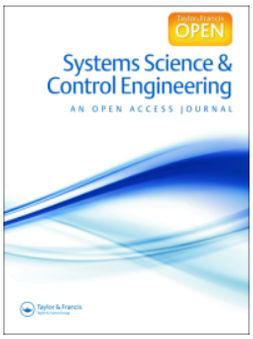

Taylor & Francis
Taylor & Francis Group

Systems Science & Control Engineering

An Open Access Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tssc20

Lightweight human activity recognition method based on the MobileHARC model

Xingyu Gong, Xinyang Zhang & Na Li

To cite this article: Xingyu Gong, Xinyang Zhang & Na Li (2024) Lightweight human activity recognition method based on the MobileHARC model, Systems Science & Control Engineering, 12:1, 2328549, DOI: 10.1080/21642583.2024.2328549

To link to this article: https://doi.org/10.1080/21642583.2024.2328549

Published online: 26 Mar 2024.

SYSTEMS SCIENCE & CONTROL ENGINEERING: AN OPEN ACCESS JOURNAL
2024, VOL. 12, NO. 1, 2328549
https://doi.org/10.1080/21642583.2024.2328549

Taylor & Francis
Taylor & Francis Group

Lightweight human activity recognition method based on the MobileHARC model

Xingyu Gong, Xinyang Zhang and Na Li

Department of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an City, Shaanxi Province, People's Republic of China

ABSTRACT

In recent years, Human activity recognition (HAR) based on wearable devices has been widely applied in health applications and other fields. Currently, most HAR models are based on the Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), or their combination. Recently, there have been proposals based on Transformer and its variant models. However, due to the fact that these models have sequential network structures and are unable to simultaneously focus on local and global features, thus, resulting in a reduction in recognition performance. In addition, along with the substantial computational resources required by Transformers, they are not suitable for resource-constrained devices. In this paper, the primary distinction of our proposed model from other hybrid models that combine CNN and Transformer is that our model adopts a completely new parallel network architecture and primarily focuses on lightweight design. Particularly, We proposed the Mobile Human Activity Recognition Conformer (MobileHARC), which adopts the parallel structure with a lightweight Transformer and CNN as the backbone networks. Furthermore, we proposed the Inverted Residual Lightweight Convolution Block and Multiscale Lightweight Multi-Head Self-Attention Mechanism. We systematically evaluate the proposed models on four public datasets. Experimental results show that MobileHARC achieves superior recognition performance, and uses fewer Floating-Point Operations per Second (FLOPs) and parameters compared to current models.

ARTICLE HISTORY
Received 21 November 2023
Accepted 4 March 2024

KEYWORDS
Human activity recognition; sensors; lightweight model; transformer

1. Introduction

Human activity recognition (HAR) refers to data collection, modelling, and recognition. It plays a significant role in various scenarios such as health care (Kalantarian et al., 2016), smart homes (Kyle & Diane, 2014), security monitoring (Niu et al., 2004), human-computer interaction (Kim & Yu, 2012), and emotion recognition (Fortino et al., 2015). HAR systems utilize sensors embedded in smartphones and wearable devices to capture human activity data and recognize daily human activities (Dalin et al., 2022). In recent years, an increasing number of researchers have been exploring the use of Deep learning algorithms to tackle the HAR problem. Deep Neural Network (DNN) -based research has become increasingly popular. In deep learning models, the feature extraction process is performed automatically, and the resulting performance is excellent. Therefore, several DNN-based HAR studies have been conducted (Demrozi et al., 2020; Khan & Ghani, 2021; Wang et al., 2019).

Currently, HAR models are mainly based on CNN, Recurrent Neural Network (RNN), or their combination. The LSTM and gated recurrent units (GRUs), which are types of RNN series models, feature a gate structure that can memorize the long-term state of the previous input and learn the sequential context of time-series data (Chen et al., 2017; Zhao et al., 2018a). The CNNs sequence models are widely used in the field of imaging and are also frequently applied to time series data. CNN models can accurately extract spatial features from spectrograms transformed from temporal sequences of data or multivariate time series data. Compared to RNNs, CNNs offer the advantage of high computational efficiency. The CNN subnet is used to extract local features of different sensor data through shared filters (Fakhruldin et al., 2017; Yang et al., 2015). With the research on HAR models progressing, the hybrid deep learning models composed of CNNs and RNNs outperform conventional CNN and RNN models to achieve excellent recognition performance (Edel & Köppe, 2016; Mekruksavanich & Jitpattanakul, 2020; Mukherjee et al., 2020; Ordóñez & Roggen, 2016; Qian et al., 2019; Siraj & Ahad, 2020). While hybrid networks achieve excellent recognition performance, however, unidirectional LSTM or bidirectional LSTM, have two limitations. Firstly, they must predict the

CONTACT Xinyang Zhang 18409313430@163.com, 2855590549@qq.com

output at the next time step based on the output of the previous time step, which cannot be processed in parallel. Secondly, while the introduction of LSTM partially addressed the vanishing gradient problem in RNN, due to the limited memory capacity of LSTM memory units, it still tends to forget information from a long time ago. The introduction of the Attentional model applied attention mechanisms to the field of activity recognition, enabling models to have stronger representational power when dealing with sequences and effectively solving the problem of non-parallel computation in LSTM. The Attention mechanisms have been incorporated into hybrid models to improve recognition performance by allowing neural networks to focus on more important information (Alireza et al., 2021; Vishvak & Thomas, 2018). The commonly used temporal attention mechanism does not effectively tackle the 'forgetting' defect of RNN, limiting the ability of existing hybrid models to capture long-term information (Shengzhong et al., 2020).

Recently, the Transformer model's self-attention mechanism, which excels at attending to global features, has made a significant impact on the natural language processing domain. The transformer was first proposed in the field of speech recognition, it offers the advantage of better computational efficiency than RNN-series models and good extraction of long-term dependency information (Vaswani et al., 2017). In addition, Transformer has also been introduced in the field of computer vision, and it has achieved excellent experimental results (Dosovitskiy et al., 2020). In HAR models based on Transformer and its variants have been proposed, including the pure Transformer (Dirgová et al., 2022; Saif et al., 2020). The combination of RNN and Transformer (Carlos et al., 2020) as well as the ConvTransformer network was first applied for HAR (Zhang, 2022; Zhang et al., 2023). The Conformer involves adding CNN layers within the Transformer to enhance the capability of local feature extraction (Kim et al., 2022). The various variants of the Transformer mentioned above aim to enhance the local feature extraction capability of a Transformer. However, these models do not individually extract and effectively integrate local and global features. Regarding Transformer-based models and their variants, many studies have overlooked the computational cost and parameters, there is a significant body of research on lightweight models based on Transformers in the fields of computer vision and natural language processing. with only a few focussing on the efficiency of Transformer models in the HAR field.

As mentioned above, How to strike a balance between the recognition performance of a HAR model and the model's parameter and computational complexity based on existing methods? How to propose a lightweight HAR model that ensures both model efficiency and effective activity recognition? Based on the current challenges in the field of HAR, we analyse existing research and propose improvements from the following perspectives. From the study of the CNN-LSTM model, we find that it is due to the limitations of convolutional and recurrent neural networks. Therefore, better performance is achieved with a hybrid model or a hybrid model combined with an attention mechanism. Replacing the original LSTM network with Transformer. CNN collects local features through convolution kernels, Transformer exhibits faster convergence and higher accuracy than LSTM. These Variant models based on the Transformer have sequential network structures and are unable to simultaneously focus on local and global features. To design a lightweight model, we proposed lightweight convolution and lightweight Transformer as the backbone networks and proposed the lightweight MobileHARC model, which adopts the parallel network structures may better extract global and local features. Therefore, Our contributions can be summarized as follows:

- In order for the model to separately focus on local and global features, We proposed a completely novel parallel network structure for our MobileHARC the hybrid model.
- Considering the lightweight design of the model, we proposed the Lightweight Convolution Block and Lightweight Multi-Head Self-Attention Mechanism for lightweight MobileHARC.
- A systematic evaluation of MobileHARC, and other models on four public datasets HOSPITAL, SKODA, UCI-HAR, and OPPORTUNITY not only in terms of performances but also in terms of FLOPs and Parameters.

The rest of the paper is organized as follows: Section 2 summarizes related work on sensor-based HAR research. Section 3 provides a detailed description of the MobileHARC architecture for HAR. Section 4, We introduce the experimental setup and analyse the experimental results. Finally, Section 5 concludes the paper with a summary of the experimental results and relevant conclusions.

2. Related work

The rapid development of Deep learning has garnered increasing attention in the field of sensor-based HAR. Although traditional machine learning algorithms based on manual, feature engineering are still being researched, they have been overshadowed by the advancements of Deep learning. For instance, In Silva and Galeazzo (2013), researchers propose the K-nearest neighbour (KNN) algorithm to classify handcrafted features extracted from sensor accelerometers. In Tian and Chen (2016), the

authors combined sequence features with wavelet coefficients and employee support vector machines (SVM) for activity classification. However, manual feature engineering is prone to human errors and limited in its ability to extract deeper-level features, which significantly impacts the classification performance of the models. In contrast, based on deep learning of HAR models achieve higher recognition accuracy compared to traditional machine learning algorithms and can be applied in various scenarios. They leverage the power of neural networks to automatically learn meaningful representations from raw sensor data, enabling the extraction of more complex and hierarchical features.

Early deep-learning models for HAR primarily focussed on using CNN and multi-branch CNNs to extract or fuse features from different sensors. Fakhrulddin et al. (2017) applied 2D-convolution and pooling operations to multiple sensor data and proposed CNNs models to reduce interference between different sensor modalities. Xu et al. propose a deformable CNN model. For different active samples, the model can adaptively adjust the receptive field and sampling location further to enhance the modelling of local information (Xu et al., 2022). Li et al. propose a two-stream CNN. The network can adaptively learn from two parallel streams, effectively reducing the computational resource consumption for feature fusion and classification of multi-scale convolution (Li et al., 2019). In Ronao et al. (2016), the authors proposed a deep CNN model and explored the impact of different hyperparameters on classification accuracy. Yang et al. proposed using deep CNNs to process multi-channel sequential signals (Yang et al., 2015). However, CNNs alone cannot capture global temporal dependencies, so the RNN (Cho et al., 2014; Hochreiter & Schmidhuber, 1997) were introduced in the HAR field. Chen WH et al. proposed using Long Short-Term Memory (LSTM) recurrent neural networks in HAR domain (Chen et al., 2017). Ullah et al. use a multi-layer LSTM for learning long-term sequences in temporal optical flow features for activity recognition (Ullah et al., 2018). Bokhari et al. use a simplified network gated recurrent unit (GRU) of LSTM with fewer parameters and showed faster convergence, but the recognition effect was not much different from LSTM (Bokhari et al., 2021). Zhao et al. propose a residual bidir-LSTM model to prevent the over-complexity in the LSTM hidden layer feature space. Residual connections and bidirectional units can ensure the effectiveness of information transfer and time dimensions (Zhao et al., 2018b). Due to the potential benefits of combining CNNs and RNNs, Ordóñez and Roggen (2016) attempted to combine CNNs and LSTMs and proposed a DeepConvLSTM model for activity recognition. Li et al. adopt a hybrid of multi-scale CNN with Bi-LSTM, which helps improve the feature extraction process and performance (Li & Wang, 2022). In addition, by combining CNN with stacked autoencoder (SAE) (Ni et al., 2020) and GRU (Dua et al., 2021), the convergence speed and feature adequacy are better than in a single model. Experimental results showed that this hybrid model outperformed previous models in recognition accuracy and improved the recognition of similar activities. To address the 'forgetfulness' issue of LSTMs, time attention mechanism was introduced based on the DeepConvLSTM model. In Vishvak and Thomas (2018), the authors applied an attention mechanism to the input sensor sequences, learning a set of weights to consider time dependencies, and then inputting them to the DeepConvLSTM model to extract local and global features. This significantly improved the classification performance on benchmark datasets. Although hybrid models have achieved good recognition results, the time attention mechanism still cannot effectively address the 'forgetfulness' problem of RNNs.

Recently, the self-attention mechanism of Transformer models has shown great success in capturing global features and has been applied in natural language processing(NLP) and in the vision domains (Aravind et al., 2021; Jacob et al., 2018; ., 2020). In the HAR field, Saif Mahmudi et al. abandoned traditional RNN structures and first applied Transformers to sensor-based HAR. The model structure mainly consists of multiple self-attention modules, multimodal attention, and global temporal attention (Vaswani et al., 2017). However, a single timestamp of recorded signals has very limited semantic information, and Transformers have a strong advantage in capturing global features but lack the ability to extract local features, making it difficult to associate a single timestamp into a complete activity. Saif et al. (2020) proposed a recognition model that combines LSTM networks and multiple self-attention modules. The LSTM network is used for local information association, and then the Transformer models are used to model global information. However, this model mainly focuses on temporal correlation and lacks the association between different sensor data. ConvTransformer (Dirgová et al., 2022) combines CNNs and Transformers for self-supervised human activity recognition, achieving better recognition results compared to unsupervised models with multiple datasets. Xiao S et al. considered that the spatial features of sensor data also affect activity recognition performance and proposed a dual-branch activity recognition model called TTN based on Transformers. One branch of the network is used to extract spatial features, while the other branch extracts temporal features, and finally, the temporal and spatial features are fused for the final activity classification (Xiao et al., 2022). In Zhang (2022) an HAR framework composed of an Inertial Measurement Unit (IMU) fuse

block and an applied ConvTransformer subnet and the features of different modalities can be aggregated more effectively. The various variants of a Transformer mentioned above aim to enhance the local feature extraction capability of a Transformer. However, these models do not individually extract and effectively integrate local and global features.

There is a significant body of research on lightweight models based on Transformers in the fields of computer vision and natural language processing. Regarding Transformer-based models and their variants, many studies have overlooked the computational cost and parameters, with only a few focussing on the efficiency of Transformer models in the HAR field. Wu et al. present an efficient mobile NLP architecture, Lite Transformer to facilitate deploying mobile NLP applications on edge devices. The key primitive is the Long-Short Range Attention (LSRA), where one group of heads specializes in the local context modelling by convolution while another group specializes in the long-distance relationship modelling by attention (Wu et al., 2020). Mehta et al. combine the strengths of CNNs and ViTs to build a light-weight and low latency network for mobile vision tasks, and introduce MobileViT, a light-weight and general-purpose vision transformer for mobile devices. MobileViT presents a different perspective for the global processing of information with transformers (Mehta & Rastegari, 2021). In the field of HAR, although many deep learning-based recognition models have been proposed, deploying HAR systems on resource-constrained devices between recognition performance, computational cost and model size requires a balance. For CNN-LSTM models, studies (Andar et al., 2018; Andrey, 2018) have attempted to use, lightweight and shallow CNNs to build recognition models, achieving good recognition performance while effectively reducing the demand for computational resources. Raza et al. propose a novel lightweight transformer, which can combine the advantages of RNNs and CNNs without their major limitations, and also TransFed, a more privacy-friendly, federated learning-based HAR classifier using lightweight transformer (Raza, 2021). EK et al. present Human Activity Recognition Transformer (HART), a lightweight, sensor-wise transformer architecture that has been specifically adapted to the domain of the IMUs embedded on mobile devices, and present evaluations across various architectures on their performances in heterogeneous environments and show that the models can better generalize on different sensing devices or on-body positions (EK et al., 2022). Xu et al. propose a series of techniques and accordingly adapt LIMU-BERT to IMU sensing tasks. The designed models are lightweight and easily deployable on mobile devices. With the representations learned via LIMU-BERT, task-specific models trained with limited labelled samples can achieve superior performances (Huatao et al., 2021).

In summary, there are still some limitations in sensor-based HAR models. Firstly, for Transformer-based models or their variants, the sequential architecture cannot independently focus on both local and global features, affecting the recognition performance of the models. Secondly, the HAR models based on Transformers are computationally heavy and not well-suited for resource-constrained devices. In this paper, regarding the issues summarized in the HAR model, we propose the lightweight MobileHARC architecture.

3. MobileHARC for wearable HAR

This section introduces we propose the Mobile Human Activity Recognition Conformer (MobileHARC). The MobileHARC is mainly composed of four components: the Inverted Residual Lightweight Convolution Block, the Lightweight Transformer block with Multiscale Lightweight Multi-Head Self-Attention Mechanism and Aggregation and prediction layer. The model structure is illustrated in Figure 1. The MobileHARC adopts a completely novel parallel network structure. Initially, the lightweight CNN layer aggregates sensor data from individual timestamps into motion segments. These segments are then separately fed into the CNN Block for extracting local features, and the Transformer module for capturing global temporal features. The feature fusion module combines the local and global features, which are subsequently inputted into the classification module for activity recognition. The following subsections describe the specific details of the model.

3.1. Inverted residual lightweight convolution block

The convolutional Block utilizes to fully extract activity local features in the MobileHARC. It starts by using 1×1 convolutional layer to increase the dimensionality of the input sensor data. Then, the data is passed through 3×1 convolutional layer for feature extraction. Once the feature extraction is complete, 1×1 convolutional layer is used to reduce the dimensionality of the features. Furthermore,we have designed a lightweight convolutional module.

The lightweight Convolution Block primarily employs Depthwise Separable Convolution (DSC) (Andrew et al., 2017), and the Inverted Residual Block. The lightweight Convolution Block is illustrated in Figure 2. DSC decomposes the standard Convolution into two parts: Depthwise Convolution and Pointwise Convolution. This decomposition significantly reduces the model's parameter count and computational complexity. In the Depthwise

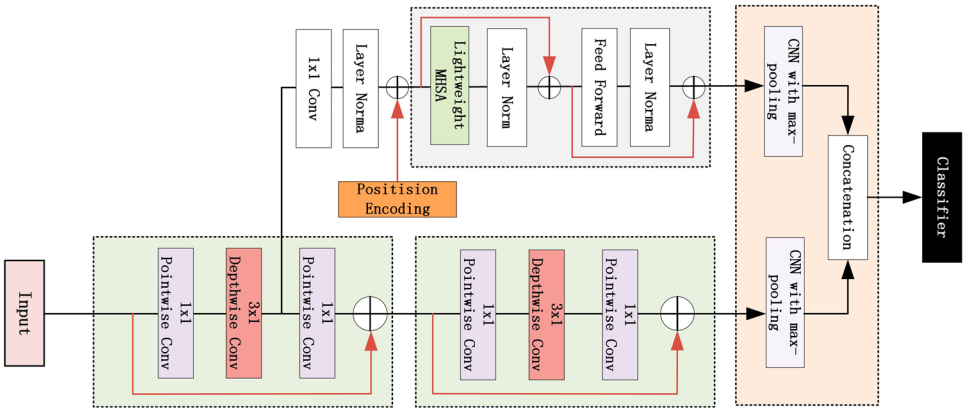

Figure 1. Overview of our proposed MobileHARC.

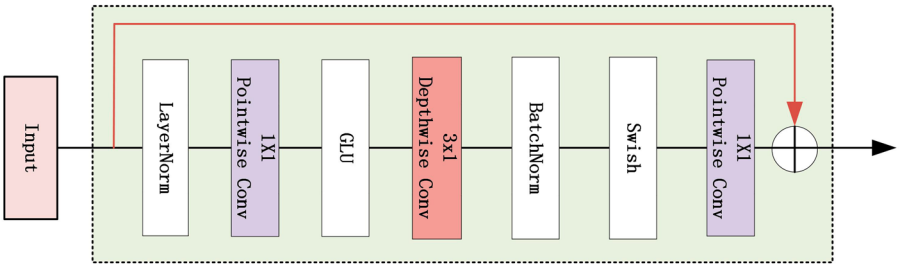

Figure 2. Inverted residual lightweight convolution block.

Convolution, each feature channel of the input is separated with a dedicated kernel. Then, the outputs of the Depthwise Convolution are combined using Pointwise Convolution, which is essentially 1 × 1 convolution. In the context of DSC, Pointwise Convolution serves two purposes: firstly, it allows DSC to freely adjust the number of output channels, and secondly, it helps to fuse the feature channels from the Depthwise Convolution output. The Depthwise Separable Convolution can be formulated as:

$$DSC(x) = SoftMax(DepthwiseConv(PW(x)) \quad (1)$$

where $DepthwiseConv(\cdot)$ represents the Depthwise convoltion layer. the $PW(\cdot)$ represents the pointwise convolution operation, which involves using 1 × 1 convolution to merge the features extracted by the Depthwise convolution.

The Depthwise convolution cannot change the number of output channels, so the input and output channel numbers are the same. However, if the input channel number is small, the Depthwise convolution can only extract features in a low-dimensional space, which may lead to insufficient feature extraction. To address this, before applying the Depthwise convolution, the input data's channel dimension is 'expanded' using a Pointwise convolution to allow the Depthwise convolution to extract features in a higher-dimensional space. Finally, the extracted features are fed into a Pointwise convolution for feature fusion and dimensionality reduction. The residual connections help preserve the original data features, thereby improving the model's performance. By combining Depthwise Separable Convolution with inverted residual structures in the lightweight convolution Block.

3.2. Multiscale lightweight multi-head self-attention mechanism

The self-attention mechanism is effective in capturing relationships between different positions in a sequence, and can process the entire sensor sequence information in parallel, effectively addressing the 'forgetting' problem of RNN networks. Additionally, it offers a more efficient way of handling the entire sensor sequence input and aids in handling long-range dependencies, enabling the model to better capture information from the distant positions in the sensor sequence. It can effectively enhance the model's performance in activity recognition. The self-attention mechanism is a method for dynamically allocating attention weights to different positions in a sequence to calculate the weighted sum of each element. It allows the model to focus on different positions of information when processing input sequences, instead of applying the same fixed-weight convolution kernel to all positions, as in traditional convolutional methods. The

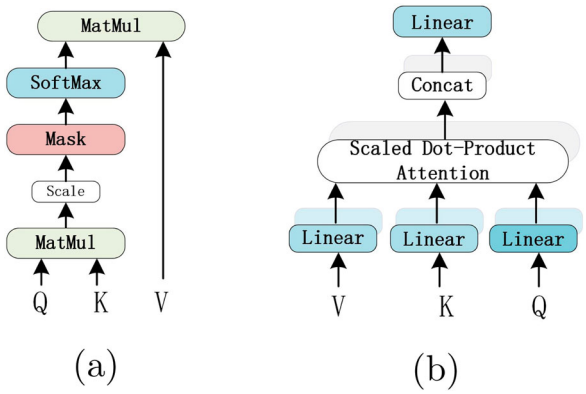

Figure 3. The structure principles of MHSA. (a) self-attention mechanism and (b) MHSA.

Multi-Head Self-Attention Mechanism (MHSA) concerts together multiple self-attention heads and then passes in a linear layer to get the final output. The self-attention mechanism and MHSA is shown in Figure 3. For a given position's element, attention weights are obtained by calculating the correlation or similarity with elements at other positions. This is typically accomplished using dot products or other similarity measures. Using the calculated weights, a weighted sum is computed for each element in the sequence. This enhances the model's focus on important elements, thereby improving the model's performance.

The Self-attention Receiving the input $X \in R^{L\times d_{in}}$ or the output of the previous. Moreover, involves three distinct linear transformations to obtain the query(Q), key(K), and value(V) representations is defined as:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2}$$

where $W_Q \in R^{d_{in}\times d_k}$, $W_K \in R^{d_{in}\times d_k}$, and $W_V \in R^{d_{in}\times d_v}$ are projection parameters. Then, the attention weights are calculated by the similarity between queries and keys, and finally the attention weights are applied to the values to output a weighted sum. The Self-Attention Mechanism is defined as:

$$f_{sa}^{(h_j)} = Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q,K,V are learned linear transformations for the input, and $Q \in R^{L\times d_k}$, $K \in R^{L\times d_k}$ and $V \in R^{L\times d_v}$, and $\sqrt{d_k}$ is a scaling factor, is used for normalization.

The d_k is the Q, K dimension. Attention weights are obtained by applying $Softmax(\cdot)$ function on the scaled dot products of queries and keys, and then perform a weighted calculation with matrix V to obtain the output. The Transformer block consists of two main components: the MHSA and the Feed-Forward Neural layer. Since the Transformer does not use the recurrent structure, it cannot exploit the order information of sensor input. Therefore, the Transformer needs to use the positional encoding to save the relative or absolute position of the sample in the sequence. Finally, MHSA concerts together multiple Self-Attention heads and then passes in a linear layer. The MHSA as follows:

$$H_{mhsa} = [f_{sa}^{(h_1)}, \ldots, f_{sa}^{(h_n)}]W_O \tag{4}$$

where $W_O \in R^{nd_v\times d_o}$ represents the parameter matrix of the linear layer.

Before the feed-forward neural network, layer normalization and residual connections are added to the structure. This design accelerates the model convergence. The output is then passed into the feed-forward layer with the same operations.

The Multi-Head Self-Attention Mechanism is computationally heavy in terms of computational resources. To reduce the model's computational burden, a combination of Dynamic convolution and Self-Attention Mechanism can achieve a lightweight attention mechanism. Dynamic convolution (Felix et al., 2019) is a form of lightweight convolution where multiple parallel convolution kernels are dynamically aggregated based on attention scores depending on the input. Since the size of the convolution kernels is small, combining multiple kernels is computationally efficient, and they are aggregated based on attention scores, which provides stronger feature representation capabilities. The Dynamic convolution can be formulated as:

$$\begin{aligned} O_{DynamicConv} &= DynamicConv(X,i,c) \\ &= DepthwiseConv(X, f(x_i)_h, :, i, c) \end{aligned} \tag{5}$$

where $f(x_i) = \sum_{c=1}^{d} W_{h,j,c}^{Q} X_{i,c}$, and W_Q is a linear weights.

The lightweight attention mechanism in the MobileHARC is designed with a dual-branch approach. The lightweight attention mechanism is illustrated in Figure 4. The input sensor data is split into two parts along the channel dimension, which are then separately fed into the self-attention mechanism and the dynamic convolution branches. Finally, the features from both branches are merged. By reducing the data's channel dimension when entering the multi-head self-attention mechanism, this approach significantly reduces the model's computational complexity. The combination of the multi-head self-attention mechanism and the dynamic convolution module in parallel placement allows dynamic convolution to non-linearly aggregate multiple sizes of convolution kernels based on attention, facilitating multi-scale feature extraction. On the other hand, the multi-head self-attention mechanism can effectively focus on global temporal features. The merging of the features from both

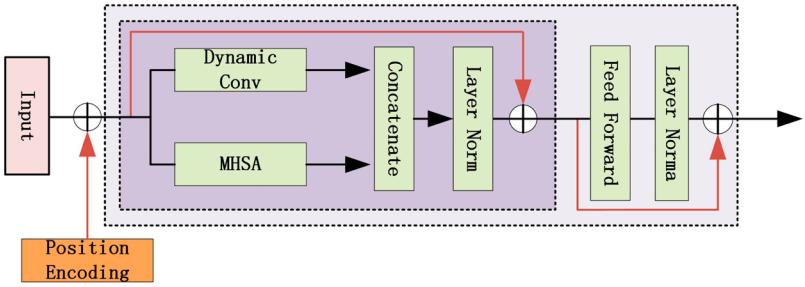

Figure 4. Multiscale lightweight attention mechanism for transformer.

branches leads to a more powerful feature representation capability. The lightweight attention mechanism can be formulated as:

$$O_h = Concat(H_{mhsa}, O_{DynamicConv}) \tag{6}$$

where H_{mhsa} represents the multi-head self-attention mechanism, obtained lightweight attention output matrix O_h is used for the Add & Norm layer, and finally input Feed-Forward layer.

3.3. Aggregation and prediction layer

In order to adapt to the input of the prediction layer, the local and global features extracted by the feature extraction layers need to be fused. The Aggregation layer is composed of convolutional layers and max-pooling operations:

$$\begin{aligned} u^x &= v(W^X, X) \\ &= ReLU(Pooling(Conv(W^X; X))) \end{aligned} \tag{7}$$

where $v(\cdot)$ represents four operations: convolutional layer, one-dimensional max-pooling, dropout layer, and the ReLU activation function.

The convolutional layer has a kernel size of w and generates l convolutional feature maps. The one-dimensional max-pooling is used to select the maximum value from specific convolutional feature maps, capturing the most important features within the convolutional feature maps. The feature fusion layer utilizes these operations to combine and refine the local and global features before passing them to the classification layer. Finally, the features after convolution and pooling are merged, and the feature vector U is generated by concatenating the two output vectors:

$$U = [u^L; u^G] \tag{8}$$

The classification layer of the MobileHARC model consists of a fully connected layer followed by the softmax function. The loss function used is the standard cross-entropy (CE-loss) loss function.

4. Experimental evaluation

This section introduces the datasets used in the experiments, describes the evaluation metrics used to measure the model's performance, and provides details about the experimental settings, including data preprocessing, model implementation, and training process.

4.1. Experimental setup

4.1.1. Datasets

This paper selects four widely used public datasets, namely SKODA, HOSPITAL, UCI-HAR and OPPORTUNITY. The selected datasets encompass multiple categories of human daily activities, making them suitable for evaluating the performance of HAR models. Detailed information about the chosen datasets is summarized in Table 1.

- **SKODA:** The SKODA dataset (Thomas et al., 2008) is a specialized dataset used to track worker activities in production line scenarios. It captures the activities of one assembly line worker performing 10 different operation gestures along with a null class in the context of automobile production. Accelerometers were placed on both the left and right arms of the participant, and the sensors were sampled at a frequency of approximately 98 Hz.
- **HOSPITAL:** The HOSPITAL dataset (Artur et al., 2018) was collected from 12 hospitalized elderly patients. In this dataset, inertial sensors were placed on the clothing of the elderly patients. It includes data from 7 classes of daily activities. The sensors were sampled at a frequency of 10 Hz.
- **UCI-HAR:** The UCI-HAR dataset (Rui et al., 2018) was collected using the accelerometer and gyroscope of smartphones to capture activity data from 30 participants. It includes data from 6 classes of daily activities. The sensors were sampled at a frequency of 50Hz.
- **OPPORTUNITY:** The OPPORTUNITY dataset (Ricardo et al., 2013) was collected from 4 participants. All the participants wore multiple body-worn sensors

Table 1. Summary of the datasets used for our experiment.

Datasets	Subjects	Activities	Sensors
SKODA	1	11	20
HOSPITAL	12	7	2
UCI-HAR	30	6	2
OPPORTUNITY	4	18	3

and performed naturalistic kitchen routines. The performed activities include 17 sporadic gestures as well as a Null class. The signals of the triaxial accelerometer, gyroscope, magnetometer and some other internal sensors were recorded at a constant rate of 30 Hz.

In this paper, the SKODA, HOSPITAL and OPPORTUNITY datasets were segmented into multiple data windows using a sliding window approach with a window size of 24 samples and an overlap rate of 50%. For the SKODA dataset, the data collected from 60 sensor channels on the right arm using 20 accelerometers were split into an 8:1:1 ratio, with 10% for the validation set and test set, and the remaining data was used for the training set. For the HOSPITAL dataset, the data collected from the first 8 elderly patients were used for model training, 3 patients for the validation set, and the rest of the test set. For OPPORTUNITY, we use runs 4 and 5 from subjects 2 and 3 as the holdout test-set, run 2 from participant 1 as the validation-set, and the remaining data as the training-set. As for the UCI-HAR dataset, the provider has already preprocessed it. The dataset is divided into sliding windows of size 128 samples with an overlap rate of 50%. It is split into two parts, with 70% used in the training set and 30% for the test set.

4.1.2. Evaluation metrics

The performance evaluation of deep learning models depends on various metrics. In this paper, we use the following evaluation metrics: Accuracy, Precision, Recall, and the Macro F1 score (mF1) as A comprehensive evaluation score for the model. It involves calculating the precision and recall for each class and then taking the average F1-score across all classes. The weighted F1 (mWF1) extends mF1 by weighting different classes according to their sample proportion. These metrics are very commonly used. In order to measure the model's lightweight for MobileHARC, We use the Parameters and FLOPs evaluation metrics.

4.1.3. Parameter setup and experimental support

The experimental environment includes a system with the NVIDIA RTX-2080Ti GPU. All comparative experiments were implemented using the PyTorch framework, and the development environment was PyCharm. The models use the standard cross-entropy loss function to calculate the loss between the classification outputs and the true data distribution. The Adam optimizer is employed as the optimization algorithm, with an initial learning rate of 0.001 to accelerate model convergence. To prevent overfitting during the training process, dropout layers with a dropout rate of 0.5 were added to the network architecture. The maximum iteration number of training is 50. We list all the model parameters of our model in Table 2.

Table 2. List of our model hyper-parameters.

Parameters	Values
Batch size	128
learning rate	0.001
Transformer layer	2
CNN layer	2
kernel size	3x1
Dropout	0.5
d_{model}	128
head	8
Optimizer	Adam
Epoch	50

4.1.4. Comparisons of different methods

The proposed MobileHARC is compared with this method, which are listed as follows.

- **DeepConvLSTM:** This Model (Ordóñez & Roggen, 2016) is considered as a baseline model, which consists of 4 convolutional layers and 2 LSTM layers. The CNN is used for extracting local features and fusing features from multiple sensors, and then the LSTM layers are employed to capture temporal correlations in the data.
- **DeepConvLSTM_Attn:** To address the drawback of 'forgetting' in LSTM, Model (Vishvak & Thomas, 2018) enhances the DeepConvLSTM model with the use of a temporal attention mechanism. This mechanism helps the model to focus more on important temporal information and improves its performance in handling long-term dependencies in the data.
- **DDNN:** This Model (Qian et al., 2019) is a parallel network architecture consisting of three parallel branches. These branches are as follows: CNN Network Branch: This branch is used to extract local features from the data. LSTM Network Branch: This branch focuses on capturing temporal features of the data. AE (Autoencoder) Learning Branch: This branch is used to learn statistical features from the data.
- **Attend_Discriminate:** This Model (Alireza et al., 2021) claimed state-of-the-art (SOTA) performance on multiple wearable device-based HAR datasets. In this framework, the data is first input to a 1D-convolutional backbone for feature extraction. Then, the interactions between feature channels of each time-step are

constructed, followed by GRUs to learn time dependencies. Finally, intra-class compactness encouraged centre-loss and Mixup data augmentation are applied.

- **Transformer Encoder:** It is the first work to introduce a Transformer model to HAR (Saif et al., 2020), and it departs from the traditional RNN network structure. Instead of using RNNs, this model leverages the Transformer architecture to capture long-range dependencies and global temporal features, leading to improved performance in activity recognition tasks.
- **ConvTransformer:** This Model (Zhang et al., 2023) proposes a deep learning model, ConvTransformer, based on CNN,Transformer, and attention mechanism. The model first uses the CNN layer to model the local information of the sensor time series signal, then uses a Transformer to obtain the temporal correlation of the feature sequence, adds an attention mechanism to highlight essential features, and finally completes the activity recognition through the fully connected layer in activity recognition.
- **Conformer:** This Model (Kim et al., 2022) from the speech recognition domain into the HAR domain, the data is initially fed into a 1-D convolutional for aggregating local features. Subsequently, it is passed through Conformer blocks that combine convolution neural networks and transformers to model both local and global dependencies of the Conformer.

4.2. Result

Below are the experimental results, We present the experimental evaluation of the proposed model on four public datasets, which mainly include the impact of different hyperparameters of the model's performance, the classification performance and as well as FLOPs and parameters evaluation of lightweight of the MobileHARC model. In addition, we recorded the testing time of the model on the test dataset.

4.2.1. Hyperparameter parameter optimization for the model

We conducted extensive experiments on the hyperparameters affecting the model. Under various hyperparameter settings, we analysed the activity recognition performance of the model. We utilized the mF1 score to comprehensively evaluate the model's performance in different environments. The major hyperparameters that had a significant impact on the model's performance include the size of the latent sequence embedding dimension, the number of heads in the Multiscale Lightweight Multi-Head Self-Attention Mechanism (MHSA), and the number of lightweight Transformer blocks. In addition, the batch size influenced the performance. The latent sequence embedding dimension, number of MHSA heads, number of lightweight Transformer blocks, and batch size were changed under optimal hyperparameter conditions in Figure 5.

For the HOSPITAL and SKODA datasets, as the batch size increases, the model's performance improves, with the optimal performance observed at a batch size of 128. However, for the OPPORTUNITY dataset, the model's performance decreases as the batch size increases, with the optimal performance observed at a batch size of 16. For the UCI-HAR dataset, there is no significant improvement in model performance with increasing batch size; it remains relatively stable. Therefore, the optimal batch size is considered to be 128. When the latent sequence embedding dimension is set to 128, the model achieves the best performance across all datasets. Only for the OPPORTUNITY dataset, the optimal performance is observed at a latent sequence embedding dimension of 256. Nevertheless, as the latent sequence embedding dimension increases, the model's performance shows a decreasing trend. Hence, the optimal latent embedding dimension is determined to be 128. Regarding the number of MHSA (Multi-Head Self-Attention) heads, the model achieves optimal performance with 8 and 32 heads, except for the OPPORTUNITY dataset, where the optimal performance is observed at 16 heads. Therefore, 8 heads are considered the optimal choice for MHSA. For the OPPORTUNITY and SKODA datasets, as the number of Conformer blocks increases, the model's performance improves, with the optimal performance observed at 2 and 4 blocks. For the UCI-HAR and HOSPITAL datasets, the best performance is achieved with 2 blocks. Thus, the optimal number of lightweight Transformer blocks is determined to be 2. In summary, increasing batch size and MHSA heads generally lead to improved performance. The optimal latent embedding dimension is 128, and controlling the number of lightweight Transformers blocks at 2 is the preferable choice.

4.2.2. The performance evaluation of the MoblieHARC

According to the specific experimental settings, this section compares the MobileHARC model with other recognition models on the SKODA, HOSPITAL, UCI-HAR and OPPORTUNITY datasets. The purpose is to evaluate the classification performance of the MobileHARC model. The experimental results are presented in Table 3.

From the experimental results, DeepConvLSTM, serving as a baseline model in the HAR domain, is a CNN-LSTM model. it can be observed that the DeepConvLSTM_Attn and Attend_Discriminate model, which incorporates the attention mechanism into the DeepConvLSTM, As a SOTA model, Attend_Discriminate, which incorporates attention mechanism and GRU in conjunction

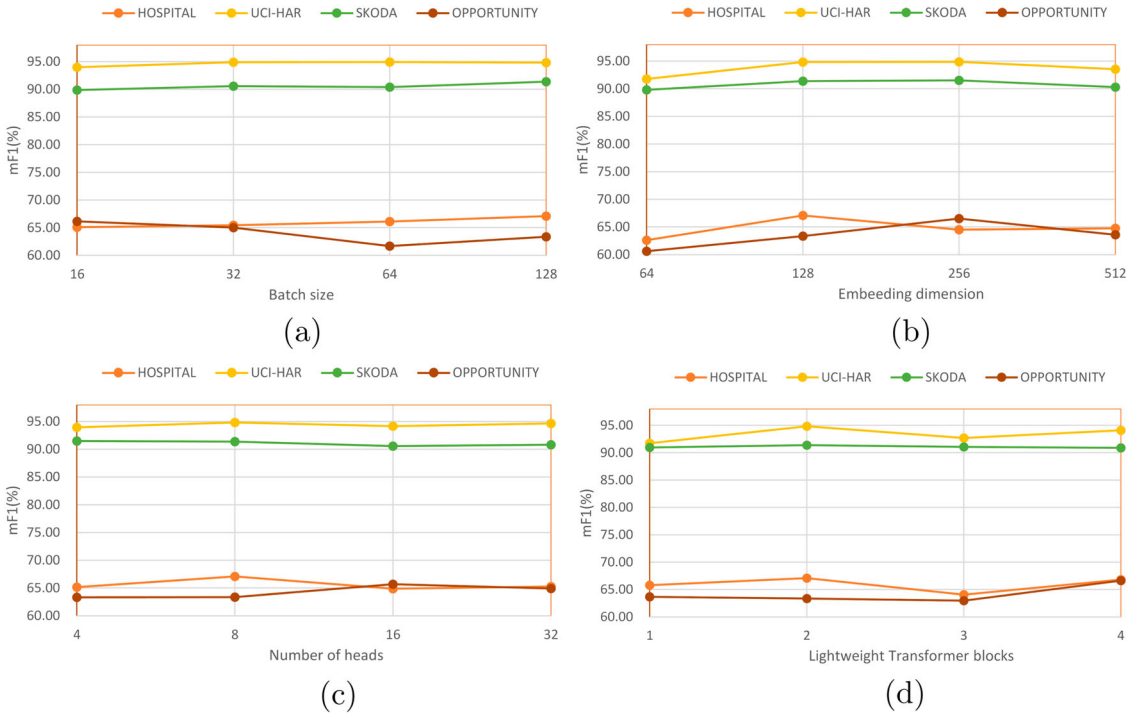

Figure 5. Optimal hyperparameters and batch size of the model. (a) mF1 according to batchsize; (b) mF1 according to latent sequence embedding dimension; (c) mF1 according to the number of MHSA; and (d) mF1 according to the number of lightweight Transformer blocks.

with 1-D convolution, exhibits a clear advantage in terms of human activity recognition performance. Outperforms the baseline model in terms of recognition precision, Accuracy, Precision, Recall and mW-F1 score, the represents an improvement of around 2% to 4%. on all dataset. DeepConvLSTM_Attn is a decrease in performance on the OPPORTUNITY dataset, particularly in terms of Precision and mW-F1 score. The overall recognition performance of the model is improved. This indicates that the attention mechanism contributes to improving the classification performance of the model. DDNN, a CNN-LSTM with added statistical features on top of time and spatial features, also shows improved recognition performance compared to the baseline model on all datasetS. It performs better than the DeepConvLSTM_Attn on the SKODA, HOSPITAL and OPPORTUNITY datasets. This suggests that enriching the feature representation of data can also enhance the model's performance, and enriching features on top of the baseline model performs better than the performance improvement achieved by the attention mechanism. However, its performance falls short of the Attend_Discriminate model on all four datasets, indicating that improving GRU and attention mechanisms is more conducive to enhancing the model's performance. Based on the analysis of the models, it can be concluded that adding an attention mechanism or increasing the variety of features can effectively enhance the recognition performance of the baseline model. However, incorporating an attention mechanism may be a better choice in terms of improving the overall recognition effectiveness.

Comparing Transformer models, the introduction of the Transformer Encoder to the HAR domain is a first. On the SKODA, HOSPITAL, and UCI-HAR datasets, the model's performance is not as good as that of DeepConvLSTM and DeepConvLSTM_Attn. However, on the OPPORTUNITY dataset, it exhibits better recognition effectiveness. The overall recognition performance of the Transformer Encoder is not as good as the CNN-LSTM model, possibly due to the Transformer's lack of ability in extracting local features, which may have diminished the model's recognition effectiveness. The ConvTransformer, which aim to enhance the local feature extraction capability of the Transformer, shows significant improvements in overall recognition performance based on the four datasets. Comparing Transformer Encoder models, the represents an improvement of around 3% to 5%. Conformer enhances local representation capability by incorporating convolution within the Transformer. On the SKODA and UCI-HAR datasets, both Accuracy and Recall outperform ConvTransformer. This indicates that augmenting local feature representation contributes to the improvement of the model's recognition performance.

Table 3. Quantitative classification performance of different methods on four public datasets.

Model	Accuracy	Precision	Recall	mW-F1
	Dataset:SKODA			
DeepConvLSTM	87.98	88.46	87.98	88.09
DeepConvLSTM_Attn	88.78	89.33	88.78	88.95
DDNN	90.49	90.78	90.49	90.58
Attend_Discriminate	89.49	90.41	89.49	89.74
Transformer_Encoder	87.16	88.07	87.16	87.35
ConvTransformer	90.68	91.23	90.68	90.88
Conformer	90.70	91.11	90.70	90.76
MobileHARC	91.32	91.75	91.32	91.46
	Dataset:HOSPITAL			
DeepConvLSTM	86.85	85.77	86.85	86.11
DeepConvLSTM_Attn	88.20	87.43	88.20	87.74
DDNN	88.64	88.52	88.64	88.51
Attend_Discriminate	90.32	89.69	90.32	89.67
Transformer_Encoder	86.35	84.54	86.35	85.18
ConvTransformer	89.63	88.60	89.63	89.04
Conformer	88.20	88.84	88.20	88.41
MobileHARC	90.73	89.73	90.73	90.07
	Dataset:UCI-HAR			
DeepConvLSTM	91.96	92.33	91.96	91.92
DeepConvLSTM_Attn	93.08	93.18	93.08	93.06
DDNN	92.53	92.69	92.53	92.44
Attend_Discriminate	94.08	94.29	94.08	94.07
Transformer_Encoder	86.70	87.27	86.70	86.62
ConvTransformer	91.82	91.90	91.82	91.81
Conformer	91.92	91.99	91.92	91.94
MobileHARC	94.88	94.97	94.88	94.87
	Dataset:OPPORTUNITY			
DeepConvLSTM	86.91	85.21	86.91	85.05
DeepConvLSTM_Attn	86.25	83.99	86.25	84.12
DDNN	89.42	89.51	89.42	89.05
Attend_Discriminate	90.46	90.01	90.46	90.06
Transformer_Encoder	88.57	88.19	88.57	88.10
ConvTransformer	89.73	89.89	89.73	89.35
Conformer	88.89	89.15	88.89	88.81
MobileHARC	90.66	90.39	90.66	90.37

Notes: The mW-F1 represent the macro F-measure, and weighted F-measure scores, respectively. The four rows of each classifier represent the Accuracy(%)/Precision(%)/Recall(%)/mW-F1 scores(%).

Comparing MobileHARC models, On the SKODA, HOSPITAL, UCI-HAR and OPPORTUNITY datasets, outperforms in terms of recognition precision, Accuracy, Precision, Recall and mW-F1 score, surpassing all models. The mW-F1 scores reach 91.46%, 90.07%, 94.87% and 90.37%. This represents an improvement of around 2% to 5% compared to the DeepConvLSTM, DeepConvLSTM_Attn and DDNN. Comparing Transformer Encoder models, represents an improvement of around 2% to 3% compared to the DeepConvLSTM, DeepConvLSTM_Attn and DDNN, outperforms the ConvTransformer and Conformer in terms of mW-F1 score, the represents an improvement of around 2%. As a SOTA model, Attend_Discriminate, MobileHARC represents an improvement of around 1%. In summary, this indicates we combine CNN and Transformer as the backbone network structure that MobileHARC adopts a novel parallelized network architecture, effectively integrating both global and local features instead of solely enhancing the local feature representation of the Transformer. This significantly improves the model's recognition effectiveness.

4.2.3. The lightweight evaluation of the MoblieHARC

Based on the specific experimental settings, this section compares the MobileHARC model with other recognition models on the SKODA, HOSPITAL, UCI-HAR and OPPORTUNITY datasets. The goal is to evaluate the Parameters and FLOPs of the MobileHARC. Additionally, the time consumption of all models on the test set is recorded. The experimental results are presented in Table 4.

From the experimental results, DeepConvLSTM_Attn, with the addition of the attention mechanism compared to the baseline model DeepConvLSTM, has a minimal impact on the model's Parameters and FLOPs, which is nearly negligible. The mF1 score shows an improvement in the SKODA, UCI-HAR, and HOSPITAL datasets, or remains relatively stable. DDNN compared to the DeepConvLSTM and DeepConvLSTM_Attn model, on the SKODA, UCI-HAR and OPPORTUNITY dataset shows a substantial decrease in and FLOPs and mF1 score shows improvement. On the HOSPITAL, UCI-HAR and OPPORTUNITY dataset shows a substantial increase in Parameters. This is mainly due to the necessity of an additional feature extraction network to enrich the model's feature representation, which significantly increases the model's parameter. In the balance between lightweight and model performance, the attention mechanism might be a preferable choice, as it compresses the model's parameter count and computational complexity. As a SOTA model, Attend_Discriminate, optimizing network parameters and computational complexity has reduced both the parameter count and computational load of the model to a very low level and the best recognition performance compared to the CNN-LSTM model. The time on the test dataset for the CNN-LSTM model remains relatively consistent, with no significant trend observed across the four datasets. Neither an attention mechanism nor enriched feature representation has a significant impact on the inference time of the model.

The Transformer Encoder, ConvTransformer and Conformer show lower Parameters on the all datasets compared to the DeepConvLSTM, DeepConvLSTM_Attn and DDNN model, and show lower Parameters compared to the Attend_Discriminate on the SKODA and OPPORTUNITY datasets. Conformer shows lower Parameters FLOPs on the all datasets compared to the DeepConvLSTM, DeepConvLSTM_Attn, DDNN and Attend_Discriminate model. The hybrid model combining Transformer and CNN is well-suited for lightweight models and is particularly mobile-friendly, making it suitable for deployment on mobile devices. Transformers and their variant models

Table 4. On four publicly available datasets, we used mF1(%) scores as the performance metric to evaluate the classification performance of the models.

Model	mF1	FLOPs	Parameters	Time(s)
		Dataset:SKODA		
DeepConvLSTM	86.19	61, 966, 336	2, 238, 027	3.8062
DeepConvLSTM_Attn	88.46	61, 957, 584	2, 228, 372	3.7043
DDNN	90.60	28, 847, 596	4, 826, 981	3.6719
Attend_Discriminate	88.21	63, 524, 247	1, 699, 404	3.7911
Transformer_Encoder	86.99	10, 562, 360	410, 453	3.6928
ConvTransformer	90.31	4, 318, 208	494, 603	3.7843
Conformer	91.08	7, 356, 416	323, 659	3.9331
MobileHARC	91.37	4, 181, 152	154, 595	3.7837
		Dataset:HOSPITAL		
DeepConvLSTM	62.87	7, 647, 744	464, 455	0.4339
DeepConvLSTM_Attn	62.12	7, 642, 576	458, 384	0.4577
DDNN	61.80	17, 012, 252	3, 192, 449	0.7277
Attend_Discriminate	65.55	6, 640, 023	272, 712	0.7954
Transformer_Encoder	51.66	9, 555, 678	401, 077	0.6299
ConvTransformer	63.36	3, 622, 912	455, 943	0.7596
Conformer	61.56	7, 261, 184	307, 911	0.5992
MobileHARC	67.10	3, 202, 160	113, 471	0.8076
		Dataset:UCI-HAR		
DeepConvLSTM	91.87	120, 095, 488	641, 606	0.5002
DeepConvLSTM_Attn	92.98	120, 037, 344	569, 143	0.4632
DDNN	92.52	115, 390, 504	4, 679, 292	0.4740
Attend_Discriminate	94.14	107, 366, 223	346, 311	0.5112
Transformer_Encoder	86.49	51, 244, 295	400, 898	0.4769
ConvTransformer	91.85	27, 302, 400	496, 774	0.4456
Conformer	92.00	38, 805, 504	384, 902	0.4930
MobileHARC	94.82	17, 613, 824	114, 221	0.5323
		Dataset:OPPORTUNITY		
DeepConvLSTM	41.28	277, 034, 496	5, 641, 362	20.0324
DeepConvLSTM_Attn	35.89	277, 019, 472	5, 625, 435	19.9538
DDNN	56.43	26, 747, 800	5, 151, 200	19.7084
Attend_Discriminate	61.97	83, 258, 391	2, 167, 251	20.6773
Transformer_Encoder	56.71	12, 133, 577	418, 058	20.3192
ConvTransformer	61.89	6, 640, 128	599, 442	19.4005
Conformer	60.21	7, 407, 104	346, 386	20.1383
MobileHARC	65.60	4, 825, 920	182, 269	19.4227

Notes: Additionally, to validate the improvements in a lightweight model, we provided a comparison of each model's Parameters/FLOPs, and testing time on the test set.

take slightly more time for inference compared to CNN-LSTM models. This is likely due to the stacked Transformer blocks, which may require additional computation time.

The proposed MobileHARC in this paper, While maintaining high recognition performance with mF1, the model exhibits an absolute advantage with lower parameters and FLOPs compared to other activity recognition models. MobileHARC takes slightly more time on the test dataset compared to the CNN-LSTM model, but it remains roughly the same as the time consumed by the Transformer Encoder, ConvTransformer, and Conformer on the test dataset. It achieves a good balance between model effectiveness and lightweight methods. This is mainly because MobileHARC, which combines lightweight convolution and lightweight attention mechanisms, significantly reduces both parameters and FLOPs compared to other models. This indicates that the lightweight convolution and attention mechanisms we designed are very effective in reducing the model's parameters and FLOPs.

4.2.4. Ablation experiment on MobileHARC

As the MobileHARC model is a lightweight model, it incorporates Inverted Residual Lightweight Convolution Block and Multiscale Lightweight Multi-Head Self-Attention Mechanism, further reducing the model's Parameters and FLOPs. The effects of these two network module improvements on the model were further experimentally validated, and the results are shown in Table 5.

From the ablation experiments, The MobileHARC$_{conv+mhsa}$ Using standard Convolution and standard Multi-head Self-Attention Mechanism. On four public datasets, the model has the lowest mF1 value, and both the parameter count and computational complexity are is the highest. It can be analyzed that the improved lightweight convolutional of MobileHARC$_{lite_conv+mhsa}$ shows a slight improvement in mF1 score, approximately around 1% on the SKODA, HOSPITAL and UCI-HAR datasets, and approximately around 4% on the OPPORTUNITY, while significantly reducing the model's Parameters and FLOPs.

Table 5. We investigate the contribution of lightweight modules by conducting an ablation study.

Model	mF1	FLOPs	Parameters
Dataset:SKODA			
MobileHARC$_{conv+mhsa}$	90.09	17, 232, 448	542,353
MobileHARC$_{lite_conv+mhsa}$	91.08	12, 098, 720	480,227
MobileHARC$_{lite_mhsa+conv}$	90.22	11, 444, 800	298,897
Dataset:HOSPITAL			
MobileHARC$_{conv+mhsa}$	62.53	11, 774, 016	452,461
MobileHARC$_{lite_conv+mhsa}$	63.79	10, 021, 552	409,951
MobileHARC$_{lite_mhsa+conv}$	66.67	8, 312, 448	319,629
Dataset:UCI-HAR			
MobileHARC$_{conv+mhsa}$	93.14	60, 991, 360	437,580
MobileHARC$_{lite_conv+mhsa}$	94.01	55, 755, 392	418,957
MobileHARC$_{lite_mhsa+conv}$	94.29	46, 966, 656	325,964
Dataset:OPPORTUNITY			
MobileHARC$_{conv+mhsa}$	62.53	22, 850, 784	718,584
MobileHARC$_{lite_conv+mhsa}$	67.39	15, 332, 160	623,101
MobileHARC$_{lite_mhsa+conv}$	61.71	12, 344, 544	277,752

[a] $_{conv+mhsa}$: Using Convolution and Multi-head Self-Attention Mechanism.
[b] $_{lite\_conv+mhsa}$: Using Lightweight Convolution and Multi-head Self-Attention Mechanism.
[c] $_{lite\_mhsa+conv}$: Using Convolution and Lightweight Multi-head Self-Attention Mechanism.

As for the improved lightweight Multi-head Self-Attention Mechanism MobileHARC$_{lite_mhsa+conv}$, On SKODA and UCI-HAR datasets, there is a small range of improved in mF1 score, Furthermore, there is a significant improvement in mF1 score on HOSPITAL, approximately around 3%, however, on the OPPORTUNITY dataset, the model's performance has declined, approximately around 1%. its impact on the model's Parameters and FLOPs is evident. The ablation experiments conducted separately on the lightweight convolutional and lightweight Multi-head Self-Attention Mechanism help analyse the model's Parameters and FLOPs, and mF1 score, which effectively validate the lightweight improvements made to the MobileHARC. Overall, the results from the ablation experiments indicate that the lightweight improvements on both the convolutional and attention modules have been beneficial.

5. Conclusions

This paper proposes the lightweight MobileHARC model, which adopts the parallel structure with lightweight Transformer and CNN as the backbone networks. In order for the model to have fewer parameters and computational requirements, we proposed the Inverted Residual Lightweight Convolution Block and Multiscale Lightweight Multi-Head Self-Attention Mechanism. The model possesses the capability of extracting both local and global temporal features, effectively enhancing the classification performance of HAR models in sensor-based domains. The model's classification performance, Parameters and FLOPs are evaluated on four publicly available datasets: SKODA, HOSPITAL, UCI-HAR and OPPORTUNITY. The experiments demonstrate that the MobileHARC achieves lower FLOPs and Parameter count while outperforming baseline models and other Transformer-based variants in terms of recognition accuracy.

The model may face the following issue: Human activity datasets generally contain data from multiple sensors, and activity data collected from multiple sensors can better represent certain activity categories. Therefore, it is important to fully consider the impact of the correlations between different sensors on the model's classification. Before performing the classification, it is essential to fuse features from different sensors to fully leverage the correlation information between them, which can significantly improve the model's classification performance. In future extensions of the main results to learning-based filtering or state estimation algorithms. This implies establishing learning models based on the main results, enabling the algorithm to adapt more effectively to new data and scenarios. Such extensions may involve integrating machine learning techniques to enhance the performance of filtering or state estimation algorithms, allowing them to handle uncertainty and dynamic environments more flexibly. Through training to different data distributions and changing conditions, thereby improving the algorithm's robustness and generalization capabilities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China [grant number 62002285].

References

Alireza et al. (2021). Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*, 1–22.

Andar, A., Abdel, M., & Jalal, L. (2018). A robust deep learning approach for positionindependent smartphone-based human activity recognition. *Sensors*, *18*, (11), 3726. https://doi.org/10.3390/s18113726.

Andrew, G., Zhu, M., Meng, L., Wei, J., & Tobias, A. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1251–1258).

Andrey, D. (2018). Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, *62*, 915–922. https://doi.org/10.1016/j.asoc.2017.09.027

Aravind, S., Tsung-Yi, L., Niki, P., Jonathon, S., Pieter, A., & Ashish, V. (2021). Bottleneck transformers for visual recognition. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16519–16529).

Artur, J., Antonio, C., & Jessica, S. (2018). *Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art*. arXiv preprint, arXiv:1806.05226.

Bokhari, S. M., Sohaib, S., Khan, A. R., & Shaf, M. (2021). Dgru based human activity recognition using channel state information. *Measurement*, *167*, Article 108245. https://doi.org/10.1016/j.measurement.2020.108245

Carlos, B., Wen-Hui, C., & Chi-Wei, K. (2020). Self-attention networks for human activity recognition using wearabledevices. In *2020 IEEE International Conference on Systems* (pp. 1194–1199).

Chen, W, Baca, C., Tou, H., & Zhiwen, Y. (2017). LSTM-RNNs combined with scene information for human activity recognition. In *IEEE, International Conference on B-Health Networking, Applications and Services*. IBEE.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint, arXiv:1406.1078.

Dalin, Z., Lina, Y., Bin, G., & Zhiwen, Y. (2022). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, *54*(4), 0360–0300. https://doi.org/10.1145/3447744

Demrozi, F., Pravadelli, G., Bihorac, A., & Rashidi, P. (2020). Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access*, *8*, 210816–210836. https://doi.org/10.1109/Access.6287639

Dirgová, L. I., Kubovčík, M., & Pospíchal, J. (2022). Wearable sensor-based human activity recognition with transformer model. *Sensors*, *22*, 1911. https://doi.org/10.3390/s22051911

Dosovitskiy, A., Mostafa, D., & Georg, H. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint, arXiv:2010.11929.

Dua, N., Singh, S. N., & Semwal, V. B. (2021). Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, *103*(7), 1461–1478. https://doi.org/10.1007/s00607-021-00928-8

Edel, M., & Köppe, E. (2016). Binarized-BLSTM-RNN based human activity recognition. In *Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Alcala de Henares, Spain* (pp. 1–7).

EK, S., Portet, F., & Lalanda, P. (2022). *Lightweight transformers for human activity recognition on mobile devices*. arXiv preprint, arXiv:2209.11750.

Fakhrulddin, A., Fei, X., & Li, H. (2017). Convolutional neural networks (CNN) based human fall detection on body sensor networks (BS) sensor data. In *International Conference on Systems and Informatics (ICSAI)* (pp. 1461–1465).

Felix, W., Angela, F., Alexei, B., & Michael, A. (2019). *Pay less attention with lightweight and dynamic convolutions*. arXiv preprint, arXiv:1901.10430.

Fortino, G., Galzarano, S., Gravina, G., & Li, W. (2015). A framework for collaborative computing and multi-sensor data fusion in body sensor networks. *Information Fusion*, *22*, 50–70. https://doi.org/10.1016/j.inffus.2014.03.005

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huatao, X., Pengfei, Z., Rui, R., Mo, L., & Guobin, S. (2021). Limubert: Unleashing the potential of unlabeled data for IMU sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, SenSys '21* (Vol. 18, pp. 220–233).

Jacob, D., Ming-Wei, C., Kenton, L., & Kristina, T. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint, arXiv:1810.04805.

Kalantarian, H., Sideris, C., Mortazavi, B., Alshurafa, N., & Sarrafzadeh, M. (2016). Dynamic computation offloading for low-power wearable health monitoring systems. *IEEE Transactions on Bio-Medical Engineering*, *64*(3), 621–628. https://doi.org/10.1109/TBME.2016.2570210

Khan, N. S., & Ghani, M. S. (2021). A survey of deep learning-based models for human activity recognition. *Wireless Personal Communications*, *120*(2), 1593–1635. https://doi.org/10.1007/s11277-021-08525-w

Kim, U., & Yu, J. (2012). Physical properties of transparent perovskite oxides (ba, la)SnO_3 with high electrical mobility at room temperature. *Physical Review*, *86*, 165205. https://api.semanticscholar.org/CorpusID:119215180

Kim, Y., Cho, W., & Lee, S. (2022). Inertial-measurement-unit-based novel human activity recognition algorithm using conformer. *Sensors*, *22*(10), 3932. https://doi.org/10.3390/s22103932

Kyle, D., & Diane, J. (2014). Heterogeneous transfer learning for activity recognition using heuristic search techniques. *International Journal of Pervasive Computing and Communications*, *14*(1). https://doi.org/10.1145/3552434

Li, X., Ding, M., & Pižurica, A. (2019). Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification. *IEEE Transactions on Instrumentation and Measurement*, *58*(4), 2615–2629.

Li, Y., & Wang, L. (2022). Human activity recognition based on residual network and bilstm. *Sensors*, *22*(2), 635. https://doi.org/10.3390/s22020635

Mehta, S., & Rastegari, M. (2021). *Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer*. arXiv preprint, arXiv:2110.02178.

Mekruksavanich, S., & Jitpattanakul, A. (2020). Smartwatch-based human activity recognition using hybrid lstm network. In *IEEE. New York, NY, USA* (pp. 1–4).

Mukherjee, D., Mondal, R., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2020). EnsemConvNet: A deep learning approach for human activity recognition using smartphone sensors for healthcare applications. *Multimedia Tools and Applications*, *79*(41-42), 31663–31690. https://doi.org/10.1007/s11042-020-09537-7

Ni, Q., Fan, Z., Zhang, L., Nugent, C. D., Cleland, I., Zhang, Y., & Zhou, N. (2020). Leveraging wearable sensors for human daily activity recognition with stacked denoising autoencoders. *Sensors*, *20*(18), 5114. https://doi.org/10.3390/s20185114

Niu, W., Long, J., Han, D., & Wang, Y. (2004). Human activity detection and recognition for video surveillance. In *IEEE International Conference on Multimedia and Expo (ICME)* (Vol. 86, pp. 719–722).

Ordóñez, F., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*(1), 115–139. https://doi.org/10.3390/s16010115

Qian, H., Pan, S. J., & Da, B. (2019). A novel distribution-embedded neural network for sensor-based activity recognition. *IJCAI-19*, 5614–5620. https://doi.org/10.24963/ijcai.2019/779

Raza, A. (2021). *Lightweight transformer in federated setting for human activity recognition*. arXiv preprint, arXiv:2110.00244.

Ricardo, C., Hesam, S., & Alberto, C. (2013). The opportunity challenge: A benchmark database for on-body sensor-based. *Pattern Recognition Letters*, *34*(15), 2033–2042. https://doi.org/10.1016/j.patrec.2012.12.014

Ronao, A., Cho, S.-B., & Qinfeng, S. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, *59*, 235–244. https://doi.org/10.1016/j.eswa.2016.04.032

Rui, Y., Guosheng, L., Qinfeng, S., & Damith, C. (2018). Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognition*, *78*, 252–266. https://doi.org/10.1016/j.patcog.2017.12.024

Saif, M., Tonmoy, M., Kishor, K., Rahman, A., Mohammad, S., Asif, C., & Amin, A. (2020). Sensor data using self-attention. In *24th European Conference on Artificial Intelligence (ECAI)* (pp. 1332–1339).

Shengzhong, L., Shuocha, Y., Jinyang, L., Dongxin, L., Tianshi, W., Huajie, S., & Tarek, A. (2020). GlobalFusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile*, *4*(1). https://doi.org/DOI:10.1145/3380999

Silva, F., & Galeazzo, E. (2013). Accelerometer based intelligent system for human movement recognition. In *IEEE International Workshop on Advances in Sensors & Interfaces* (pp. 20–24).

Siraj, M. S., & Ahad, M. A. R. (2020). A hybrid deep learning framework using CNN and GRU-based RNN for recognition of pairwise similar activities. In *Proceedings of the 2020 Joint 9th International Conference on Informatics, Electronics Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision Pattern Recognition (icIVPR)* (pp. 26–29).

Thomas, S., Daniel, R., Georg, O., Paul, L., & Gerhard, T. (2008). Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing*, 42–50. https://doi.org/DOI:10.1109/MPRV.2008.40

Tian, Y., & Chen, W. (2016). MEMS-based human activity recognition using smartphone. In *Control Conference IEEE* (pp. 3984–3989).

Ullah, A., Muhammad, K., Del Ser, J., & Baik, S. W. (2018). Activity recognition using temporal optical fow convolutional features and multilayer lstm. *IEEE Transactions on Industrial Electronics*, *66*(12), 9692–9702. https://doi.org/10.1109/TIE.41

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA* (Vol. 3058).

Vishvak, S., & Thomas, P. (2018). On attention models for human activity recognition. In *Proceedings of the ACM International Symposium on Wearable Computers* (pp. 100–103).

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11. https://doi.org/10.1016/j.patrec.2018.02.010

Wu, Z., Liu, Z., & Lin, J. (2020). *Lite transformer with long-short range attention*. arXiv preprint, arXiv:2004.11886.

Xiao, S., Wang, S., & Huang, Z. (2022). Two-stream transformer network for sensor-based human activity recognition. *Neurocomputing*, *512*, 253–268. https://doi.org/10.1016/j.neucom.2022.09.099

Xu, S., Zhang, L., Huang, W., & Wu, H. (2022). Deformable convolutional networks for multimodal human activity recognition using wearable sensors. *IEEE Transactions on Instrumentation and Measurement*, *7*, 42–50. https://doi.org/DOI:10.1109/MPRV.2008.40

Yang, J., Nguyen, M., & San, P. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proc. IJCAI. AAAI Press*.

Zhang, Y. (2022). IF-ConvTransformer: A framework for human activity recognition using IMU fusion and ConvTransformer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*(2), 1–26. https://doi.org/DOI:10.1145/3534584

Zhang, Z., Wang, W., & An, A. (2023). A human activity recognition method using wearable sensors based on convtransformer model. *Evolving Systems*, 1–17. https://doi.org/DOI:10.1007/s12530-022-09480-y

Zhao, Y., Yang, R., & Chevalier, G. (2018a). Deep residual Bidir-LSTM for human activity recognition using wearable sensors. *Mathematical Problems in Engineering*, 1–13. https://doi.org/DOI:10.1155/2018/7316954

Zhao, Y., Yang, R., Chevalier, G., Xu, X., & Zhang, Z. (2018b). Deep residual bidir-LSTM for human activity recognition using wearable sensors. *Mathematical Problems in Engineering*, *2018*. https://doi.org/10.1155/2018/7316954

Zhouyong, L., Shun, L., Wubin, L., Jingben, L., Yufan, W., Chunguo, L., & Luxi, Y. (2020). *ConvTransformer: A convolutional transformer network for video frame synthesis*. arXiv preprint, arXiv:2011.10185.