IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 August 2023, accepted 28 August 2023, date of publication 30 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3310269

APPLIED RESEARCH

Attention-Based Residual BiLSTM Networks for Human Activity Recognition

JUNJIE ZHANG[1], YUANHAO LIU[2], AND HUA YUAN[3]
[1]Hubei Key Laboratory of Digital Textile Equipment, Wuhan Textile University, Wuhan 430073, China
[2]Computer and Artificial Intelligence Collage, Wuhan Textile University, Wuhan 430200, China
[3]Wuhan Textile and Apparel Digital Engineering Technology Research Center, School of Fashion, Wuhan Textile University, Wuhan 430073, China

Corresponding author: Hua Yuan (2019009@wtu.edu.cn)

ABSTRACT Human activity recognition (HAR) commonly employs wearable sensors to identify and analyze the time series data collected by them, enabling the recognition of specific actions. However, the current fusion of convolutional and recurrent neural networks in existing approaches encounters difficulties when it comes to differentiating between similar actions. To enhance the recognition accuracy of similar actions, we suggest integrating the residual structure and layer normalization into a bidirectional long short-term memory network (BLSTM). This integration enhances the network's feature extraction capabilities, introduces an attention mechanism to optimize the final feature information, and ultimately improves the accuracy and stability of activity recognition. To validate the effectiveness of our approach, we extensively tested it on three public datasets: UCI-HAR, WISDM, and KU-HAR. The results were highly encouraging, achieving remarkable overall recognition accuracies of 98.37%, 99.01%, and 97.89% for the respective datasets. The experimental results demonstrate that this method effectively enhances the recognition accuracy of similar behaviors. A codebase implementing the described framework is available at: https://github.com/lyh0625/1DCNN-ResBLSTM-Attention.

INDEX TERMS Human activity recognition, wearable sensors, attention mechanism, residual block.

I. INTRODUCTION

Human activity recognition (HAR) refers to the process of assessing and categorizing human behavior by analyzing its activities over a specific duration. Human activity recognition (HAR) has several applications in sports, geriatric health monitoring, and safety monitoring systems [1], [2] [3], [4]. It can be divided into two approaches: one focuses on recognizing human behavior using cameras and images, while the other involves identifying activities through wearable sensor devices and sensor data [5]. The advancement of microelectromechanical systems (MEMS) has led to progressive enhancements in their characteristics, including improved sensitivity and reduced power consumption. As a result, the mainstream utilization of MEMS as wearable devices, employing inertial sensing units for human behavior recognition, has become prevalent [6], [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Shuihua Wang.

Human Activity Recognition (HAR), in wearable sensing devices, involves collecting data via sensors and then processing it to identify the activities. In recent years, several machine learning methods have been employed for human activity recognition, such as random forests (RF) [8], support vector machine (SVM) [9], and k-nearest neighbor (KNN) [2]. These methods require the manual calculation of time-domain, frequency-domain, and time-frequency domain features from the collected sensor data, followed by behavioral data classification. However, manual approaches heavily depend on the expertise of the calculator and consume a significant amount of time. Deep learning solves these problems by enabling automated feature extraction. There are many researchers who study the application of convolutional neural networks (CNNs) [10] in the field of image recognition. Zeng et al. [11] employed convolutional networks for human activity recognition. They input the collected data in the form of images into the convolutional network and observed improved recognition

accuracy. Compared to machine learning methods such as support vector machine (SVM) and random forest (RF). Recurrent Neural Networks (RNNs) [12] are widely used in the fields of natural language processing and speech recognition. RNNs can extract time series features from data. As a variant of recurrent neural networks, long short-term memory (LSTM) [13] networks can capture longer time-series features. Murad and Pyun [14] conducted experiments on the UCI-HAR [15] dataset using the LSTM network and achieved an accuracy of 96.7%. Following this, Ordóñez and Roggen [16] proposed that the DeepConvLSTM network, which combines CNN and LSTM, is a significant improvement over single neural networks for human behavior recognition. The Transformer [17] network was initially applied in natural language processing and then in the field of graphics. The Transformer network has effectively resolved the limitation of parallel computation in recurrent neural networks, allowing for efficient calculations that were previously unattainable. Buffelli and Vandin [18] introduced a multi-head attention mechanism from the transformer to identify activity data, yielding excellent identification results.

In the case of collected activity data, CNNs can extract spatial features, whereas LSTM can extract time series features. Both CNNs and RNNs possess single-feature extraction capabilities, but relying solely on individual networks leads to limited accuracy in activity recognition. The Transformer's multi-head attention computation method consumes significant computational resources and underperforms with small data volumes. The DeepConvLSTM, which combines CNN and LSTM, has obvious advantages in feature extraction, but there is a problem with unstable activity recognition.

To address the limitations of the aforementioned approaches, we introduce a neural network model called the 1DCNN-ResBLSTM-Attention model. This model combines the power of 1DCNN, residual bidirectional long short-term memory (ResBLSTM), and attention mechanism. The proposed 1DCNN-ResBLSTM-Attention model utilized a one-dimensional convolutional network (1DCNN) to extract spatial features from time series data. To establish a long-distance dependence on time series data, we incorporate a bi-directional long short-term memory (BLSTM) network. Additionally, we enhance the performance of the BLSTM network by integrating residual blocks, which are referred to as ResBLSTM. These improvements contribute to the overall performance of the network. For the acquisition of final recognition information, we employ an attention mechanism to compute the weights of each result during the recursive process of the ResBLSTM network. This approach enhances the representation ability of the final recognition information, resulting in improved accuracy in activity recognition. To optimize network performance, batch normalization (BN) [19] and layer normalization (LN) [20] were introduced to the model structure. This inclusion accelerates the convergence speed of the network and reduces the time required for behavior recognition.

The main contributions of this paper are as follows:

(1) Our research presents a deep learning-based behavior recognition model called 1DCNN-ResBLSTM-Attention. This model offers an automated feature extraction capability from sensor data, making it a versatile framework for various human behavior recognition tasks.
(2) In order to improve the performance of the model, we introduce a residual structure in the BLSTM network and construct the ResBLSTM network. This addition improves the accuracy of identification. In addition, we employ an attention mechanism that optimizes the final behavioral recognition features and further improves the accuracy of the model.
(3) In our study, we constructed ResBLSTM and BLSTM networks with different numbers of stacked layers. Experimental results show the effectiveness of introducing residual structure and layer normalization in improving model performance.
(4) To evaluate the robustness and generalization ability of our model, we conducted comparative experiments on three public datasets: UCI-HAR, WISDM, and KU-HAR. The recognition accuracy achieved on these datasets was 98.37%, 99.01%, and 97.89%, respectively, and also performs well in distinguishing similar activities.

This paper is structured as follows: Section I is the introduction, providing an overview of the research topic and its significance. Section II gives an overview of current literature. In Section III, we present an overview of the three public datasets used in our experiments and outline the preprocessing steps of the data. Section IV provides a comprehensive description of the architecture and key components of the 1DCNN-ResBLSTM-Attention model. Section V demonstrates the effectiveness of ResBLSTM and offers a detailed comparison of the experimental results obtained from our proposed model across the three public datasets. Finally, Section VI concludes this study, summarizing the findings and discussing their implications.

II. RELATED WORK

The human activity recognition method for wearable sensors can be categorized into two main approaches: machine learning-based and deep learning-based methods. Early machine learning methods required the calculation of time domain and frequency domain features from activity data, which were then used for activity classification. Feng et al. [8] employed random forests to classify activities by extracting statistical features such as mean, variance, standard deviation, skewness, and kurtosis from sensor data. Jain and Kanhangad [9] used gradient histograms and Fourier descriptors based on centroid features to extract features from acceleration and angular velocity data. They utilized two classifiers, support vector machine and k-nearest neighbor,

achieving an accuracy of 97.12% on the UCI-HAR dataset. These methods focused on the global features of the activity data. However, Aşuroğlu et al. [21] took a different approach by using Local Binary Patterns (LPB) to extract the local texture features from the activity data. They used k-nearest neighbor classifiers to emphasize the importance of local features. Subsequently, Aşuroğlu et al. [22] proposed a method that combined the time and frequency characteristics of accelerometer data to the Locally Weighted Random Forest (LWRF) machine learning algorithms. This approach demonstrated outstanding performance for complex activities.

Traditional machine learning methods necessitate manual feature extraction, which is constrained by the knowledge and expertise of researchers. In contrast, deep learning offers the advantage of automatic extraction of higher-level features. Recent researches focused on four main types of sensor-based Human Activity Recognition (HAR) deep learning methods. The first type employs Convolutional Neural Networks (CNNs) for spatial feature extraction. The second type uses Recurrent Neural Networks (RNNs) to capture temporal dependencies. The third type incorporates attention structures like Transformer to focus on relevant information. The fourth type combines CNN and RNN architectures.

Convolutional neural networks (CNNs) possess the ability to automatically extract spatial features, making them highly versatile in computer vision applications. In the context of Human Activity Recognition (HAR), CNNs can be employed to process preprocessed multi-channel sensing data and extract informative features through the use of stacked convolutional kernels. Zeng et al. [11] achieved superior recognition accuracy in multi-sensor Human Activity Recognition (HAR) by using a three-axis accelerometer. They employed CNNs to independently extract high-level features from each axis. By fusing these features, their approach outperformed traditional machine learning methods in accurately identifying human activities. Ronao and Cho [23] explored multi-channel feature computation in HAR and utilized a one-dimensional convolutional neural network (CNN) optimized for time series data. This approach resulted in a remarkable recognition accuracy of 90% across six daily activities. In other variations of CNN usage for HAR, Mekruksavanich et al. [24] employed the ResNet network to successfully identify 18 activities with an impressive accuracy exceeding 93%. These findings showcase the effectiveness of CNN-based models in achieving high accuracy and robust activity recognition in diverse scenarios.

Human activity data captured from sensors is typically represented as time series, and to effectively capture the temporal relationships, RNNs are commonly employed. In the domain of HAR, various RNN variants have been utilized to process this sequential data. One prevalent approach [14] involves the use of Long Short-Term Memory (LSTM) networks. These networks excel at encoding fragments of active data, and the information is then leveraged for accurate activity identification. Ishimaru et al. [25] adopted Bidirectional LSTM (Bi-LSTM) networks to monitor reading activities. Bi-LSTM networks have the advantage of capturing information from both past and future time steps, making them well-suited for tasks requiring bidirectional temporal context. To conserve computational resources while maintaining efficient human activity recognition, Yao et al. [26] opted for Gated Recurrent Units (GRUs) networks instead of LSTMs. GRUs offer similar capabilities to LSTMs but are more computationally efficient.

The Transformer model, incorporating a multi-head attention mechanism, has demonstrated remarkable success in both Natural Language Processing (NLP) and Computer Vision (CV). It overcomes the limitations of RNNs by enabling parallel computation while extending the receptive field to the global level. Buffelli et al. [18] constructed a pure attention HAR model and utilized transfer learning for specific user identification. Shavit and Klein [27] employed Transformer's encoder to extract features from activity data, using the encoded information for activity identification. Khan and Ahmad [28] substituted the multilayer perceptron with convolutions to create a multi-head CNN. This modification addressed the overfitting issue of Transformers when applied to small human activity datasets. These variations of the Transformer model demonstrate its adaptability and effectiveness in HAR, making it a valuable tool for capturing complex patterns and improving activity recognition performance.

Feature fusion networks are commonly employed in Human Activity Recognition (HAR) to combine features from different modalities. A popular combination approach involves integrating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively capture both spatial and temporal features, which may not be fully achievable with a single feature extraction network. Karim et al. [29] proposed to combined network named LSTM-FCN, which incorporates a fully convolutional network for spatial feature extraction and an LSTM network for time series feature extraction. These two networks operate in parallel, and their outputs are merged to surpass the performance of individual CNN or RNN networks across various datasets. Similarly, Ordóñez and Roggen [16] where CNNs and LSTM networks are connected sequentially. The feature data extracted by CNNs is then fed into LSTMs, and the encoded information from LSTMs is used for activity identification.

These approaches provide effective strategies for human activity recognition based on wearable sensing devices. We adopted a feature fusion strategy and proposed a 1DCNN-ResBLSTM-Attention model, our strategy has made some progress compared with some previous work.

III. DATASET

A. DATA PRE-PROCESSING

1) NORMALIZATION

The collected data from different sensors often exhibits non-uniform scales, and the presence of outlier samples within

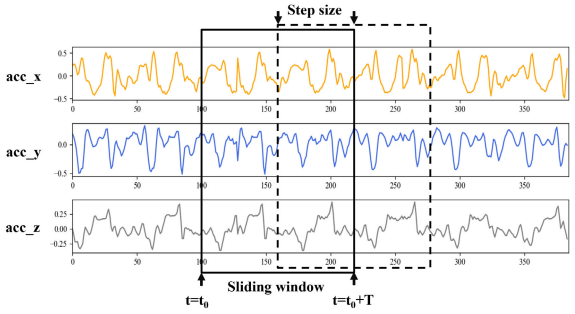


FIGURE 1. The active data is split with a sliding window with a window length of T, step_size overlapping parts of two action segments.

the data can lead to longer training times and potential convergence issues. So, it is necessary to normalize the data to a standardized range, typically [0,1]. The normalization process can be represented as follows:

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \quad (i = 1, 2, \ldots, n). \tag{1}$$

where n is the total number of samples, and x_{min} and x_{max} are the minimum and maximum values in the sample.

2) DATA SEGMENTATION

Data segmentation plays a crucial role in activity recognition, as it involves dividing the time series data into smaller segments. The length of these data fragments significantly impacts the quality and efficiency of feature extraction. The sliding window method is commonly used for data segmentation and considers two main factors: window length and window coverage. The window length typically ranges from 0.5 to 10 seconds, as reported in current research [30], [31]. If the window length is too long, it may result in a single time series containing multiple actions, making it difficult to capture the complete action. On the other hand, if the window length is too short, it may fail to reflect the holistic nature of the action. Fig.1 depicts the sliding window segmentation of three-axis accelerometer data. From top to bottom, it shows the acquisition of original sensor data for the acceleration sensor in the x, y, and z axis. The window length is denoted as T, and the window coverage is defined as $step\_size/T$.

B. DATASET DESCRIPTION

Testing the proposed model on a mobile phone behavior dataset holds significant value due to the widespread usage and diverse applications of mobile phones as wearable devices. To thoroughly evaluate the model's performance and its applicability to real-world scenarios, we utilized three distinct datasets (UCI-HAR, WISDM, KU-HAR) specifically designed for mobile phone behavior analysis. These datasets serve as comprehensive sources of information to assess the model's accuracy and effectiveness in recognizing and classifying various mobile phone behaviors.

Among the three datasets considered, both the UCI-HAR and KU-HAR datasets have undergone normalization and data segmentation. Specifically, the window length for the UCI-HAR dataset is set to $T = 128$ with a window coverage of 50%, while the window length for the KU-HAR dataset is set to $T = 300$ with a window coverage of 50%. However, the WISDM dataset has not undergone preprocessing and remains unsegmented. We follows the partitioning strategy from [32] reference and set the window length to 128 with a window coverage of 50%. For the WISDM dataset, the window length is set to 128 with a window coverage of 50%. All datasets were subjected to a train-test partitioning strategy to create separate training and testing subsets. The detailed descriptions of each dataset are as follows:

1) UCI-HAR

The UCI-HAR dataset [15] was collected from a group of 30 volunteers who wore a smartphone (SAMSUNG Galaxy S2) attached to their waist. The smartphone's three-axis accelerometer and gyroscope recorded data at a frequency of 50Hz. The dataset comprises six different types of movements, namely walking (walk), standing (stand), upstairs, downstairs, sitting (sit), and lying (lay). To prepare the dataset for analysis, the collected data were denoised using the original signal. It was then divided into segments using a sliding window approach with a window length of 128 and an overlap rate of 50%. This division resulted in a total of 10,299 data samples, each consisting of 128 data points. The dataset includes three types of data: three-axis acceleration, three-axis linear acceleration, and triaxial angular velocity. For training and testing the model, the dataset was further split into a training set and a test set. The training set accounts for 70% of the data, while the remaining 30% forms the test set. This division ensures a comprehensive evaluation of the model's performance on unseen data.

2) WISDM

The WISDM dataset [33] was collected from a group of 36 volunteers who carried a smartphone in their trouser pockets. The smartphone's three-axis accelerometer recorded data at a frequency of 20Hz. The dataset consists of six categories of activities: going upstairs (upstairs), going downstairs (downstairs), sitting (sit), standing (stand), walking (walk), and jogging (jog). To facilitate analysis, the collected data were segmented using a sliding window approach with a window length of 128 and an overlap rate of 50%. This segmentation resulted in a total of 17,158 data samples, each containing 128 data points. For training and testing the model, the dataset was further divided into a training set and a test set. The training set contains 80% of the data, while the remaining 20% forms the test set. This split ensures that the model is evaluated on unseen data and can generalize well to new instances.

3) KU-HAR

The KU-HAR dataset [34] was collected from a group of 90 volunteers who wore a waist bag containing a

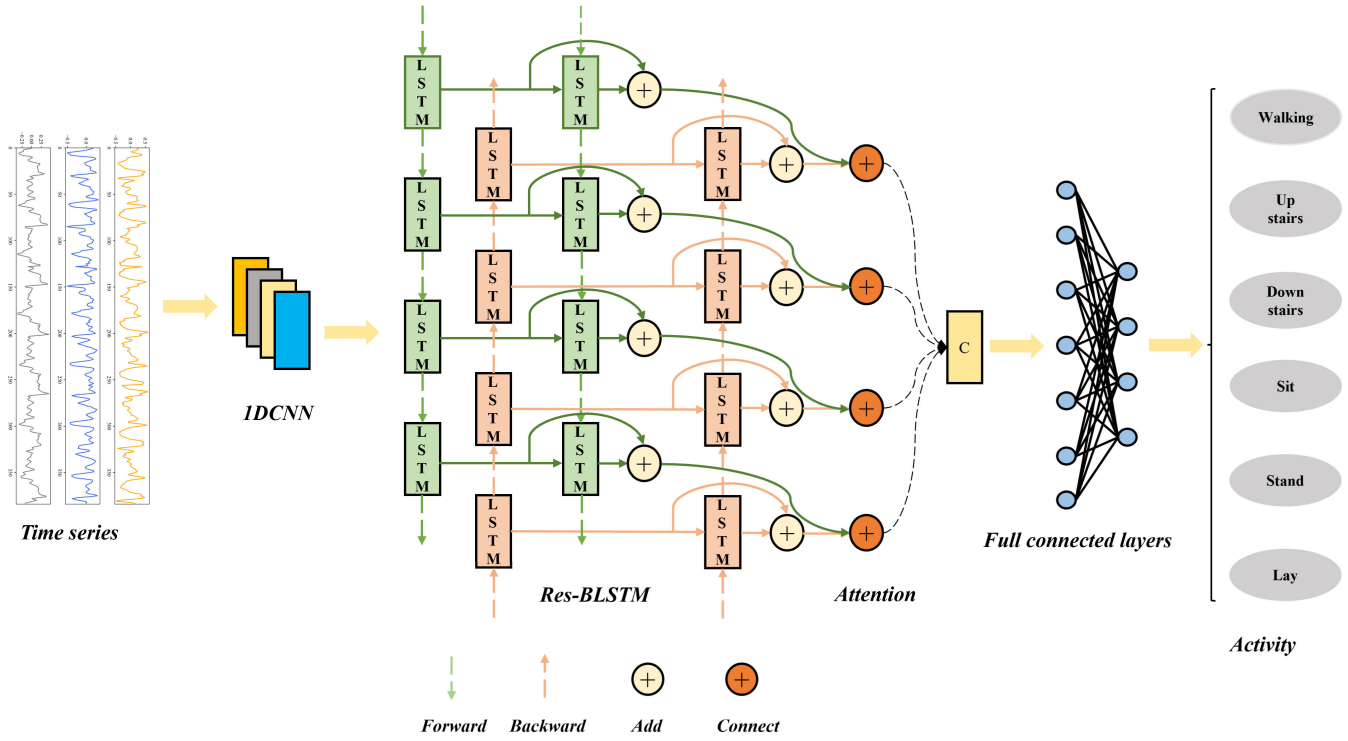


FIGURE 2. 1DCNN-ResBLSTM-Attention model structure.

smartphone during the data collection process. The smartphone's accelerometer and gyroscope recorded data at a frequency of 100Hz. The dataset consists of 18 different types of actions, including standing (std), sitting (sit), talking while sitting (talk-sit), talking while standing (talk-std), standing up from sitting (std-sit), lying down (lay), standing up from lying down (lay-std), picking up an object (pick), jumping (jump), push-ups (push-up), sit-ups (sit up), walking (walk), walking backward (walk-back), walking in a circle (walk-circle), running (run), going upstairs (up), going downstairs (down), and playing table tennis (table-tennis). Each action in the dataset has a time series length of 300 data points, corresponding to a duration of 3 seconds at a frequency of 100Hz. The data was segmented using a sliding window approach with an overlap rate of 50%. This resulted in a total of 20,750 data samples. To train and evaluate the model, the dataset was divided into a training set and a test set. The training set contains 80% of the data, while the remaining 20% forms the test set. This division allows for the model to be trained on a majority of the data and tested on unseen samples to assess its performance.

IV. PROPOSED 1DCNN-ResBLSTM-ATTENTION MODEL

Based on the human activity recognition of wearable sensing devices, activity data is first collected through Bluetooth, WiFi, radar, and other devices, and then the collected data is preprocessed, and finally, the processed data is identified. Current methods take a long time to identify and cannot distinguish between similar activities, such as going upstairs and going downstairs. To solve the problems in existing models, we propose a novel architecture called 1DCNN-ResBLSTM-Attention. The model's network structure, illustrated in Fig.2, consists of three key components: 1DCNN, ResBLSTM, and Attention. The first component, 1DCNN, is responsible for extracting spatial features from the preprocessed data. By controlling the step size of the convolution kernel, it effectively reduces the length of the time series. This enables the model to reduce recognition time. Next, the improved ResBLSTM network is used to extract time series features from the data processed by 1DCNN. By combining the strengths of bidirectional long short-term memory (BLSTM) and incorporating residual connections, the ResBLSTM component enhances the model's ability to capture long-term dependencies in the time series data. This integration boosts the model's capability to understand complex temporal patterns and improves recognition accuracy. To further optimize the final recognition features, we introduce the attention mechanism. This mode calculates weights for the feature information generated by the ResBLSTM network, allowing the model to selectively focus on the most informative parts of the input data. By emphasizing the most relevant features, the attention mechanism enhances the discriminative power of the model and improves the accuracy of activity recognition. Finally, the fully connected layer and SoftMax function are employed to classify the behavior information. The output of this classification process serves as the recognition result, providing a prediction of the specific activity being performed. In the subsequent sections, we will delve into

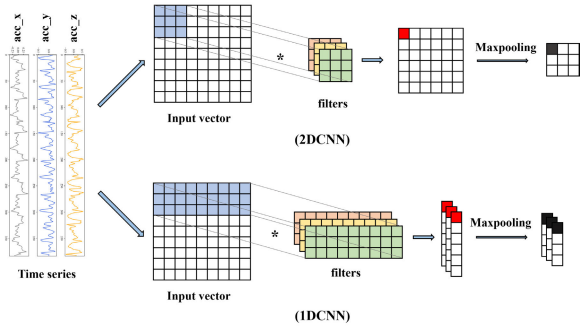


FIGURE 3. The difference between 1DCNN and 2DCNN (acc_x, acc_y, and acc_y represent the x, y, and z axes in the accelerometer, respectively).

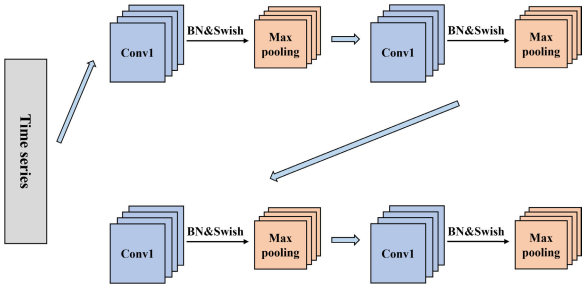


FIGURE 4. 1DCNN layer in human activity recognition structure.

a detailed explanation of each component, outlining their functionalities and contributions within our proposed model.

A. 1DCNN

CNNs have strong feature extraction capabilities in processing tensor data, making them well-suited for tasks such as image processing and human behavior recognition. In this study, we apply the one-dimensional convolutional neural network (1DCNN) [35] for effective feature extraction. As shown in Fig.3, in the acquired sensing data, the vertical axis represents the time series, and the horizontal axis represents the multi-axis channel features acquired by different sensors. The convolution of 2D convolution is local, and when the number of sensors is large, the local convolution will destroy the integrity of the sensor channels, while 1D convolution is a convolution in behavioral units, and all sensor channels are computed, so 1D convolution is chosen in the design of the model for extracting spatial features instead of using the traditional 2D convolution. In the calculation of the 1D convolution, the input data are convolved with each filter and then activated by a nonlinear activation function, which is calculated as follows:

$$X_j = f\left(\sum_{i=1}^{n}\left(W^i \cdot x_j + b^i\right)\right). \tag{2}$$

where X_j is the activated output data, W^i is the weight of the i filter, x_j represents sensing data convolved with W^i, b^i is the bias of the i filter, n is the number of filters, and f is a non-linear function.

In this paper, the convolutional layer uses the swish function [36] as the activation function. In comparison to the Relu [37] function, the swish function solves the lethality problem of the negative interval and is better suited for the sensor data. The swish function is defined as follows:

$$swish(x) = x \cdot sigmoid(x). \tag{3}$$

After the activation step, the pooling layer is utilized for downsampling. In this downsampling process, the ''same'' padding is applied. Additionally, in the 1DCNN layer, the length of time can be reduced by adjusting the stride of the pooling kernel, and the length of the time series changes as follows:

$$len_{out} = \frac{len_{input}}{s}. \tag{4}$$

where len_{input} is the input time series length, s is the pooled kernel step size, len_{out} is the pooled time series length.

During the training of neural networks, the probability distribution of input obedience in each layer undergoes continuous changes, which can lead to the issue of vanishing or exploding gradients. This phenomenon is known as the intermediate covariate shift problem [38]. In 2015, Ioffe et al [19]. introduced batch normalization (BN) as a solution to reduce the problem of intermediate covariate shifts. The fundamental concept of batch normalization involves calculating the mean and variance of a batch of data and transforming it into a new set of data with a mean of 0 and a variance of 1. By incorporating normalization into the training process, batch normalization can accelerate convergence during gradient descent, thereby reducing the training time of neural networks. The calculation process is described as follows:

$$\widehat{x_{i,k}} = \frac{x_{i,k} - \mu_k}{\sqrt{\sigma_k^2 + \varepsilon}}. \tag{5}$$

where $x_{i,k}$ is the k-dimensional component of x_i in the training set $\{x\}$, μ_k is the mean of the k-dimensional component of all samples in the training set, $\sqrt{\sigma_k^2 + \varepsilon}$ is the standard deviation of the k-dimensional component of all samples in the training set.

After the convolutional layer, the structures of Conv, BN, Swish, and Maxpooling were applied, as shown in Fig.4. In the 1DCNN section, we adopted a stacking approach to implementing this structure across four layers.

B. ResBLSTM

Human activities are inherently temporal, and relying solely on 1DCNN for spatial feature extraction is insufficient for activity recognition. The temporal sequence of the entire action must also be taken into account. RNNs exhibit favorable capabilities for processing time series data. However, as the time series grows, RNN models can suffer from gradient vanishing and information loss.

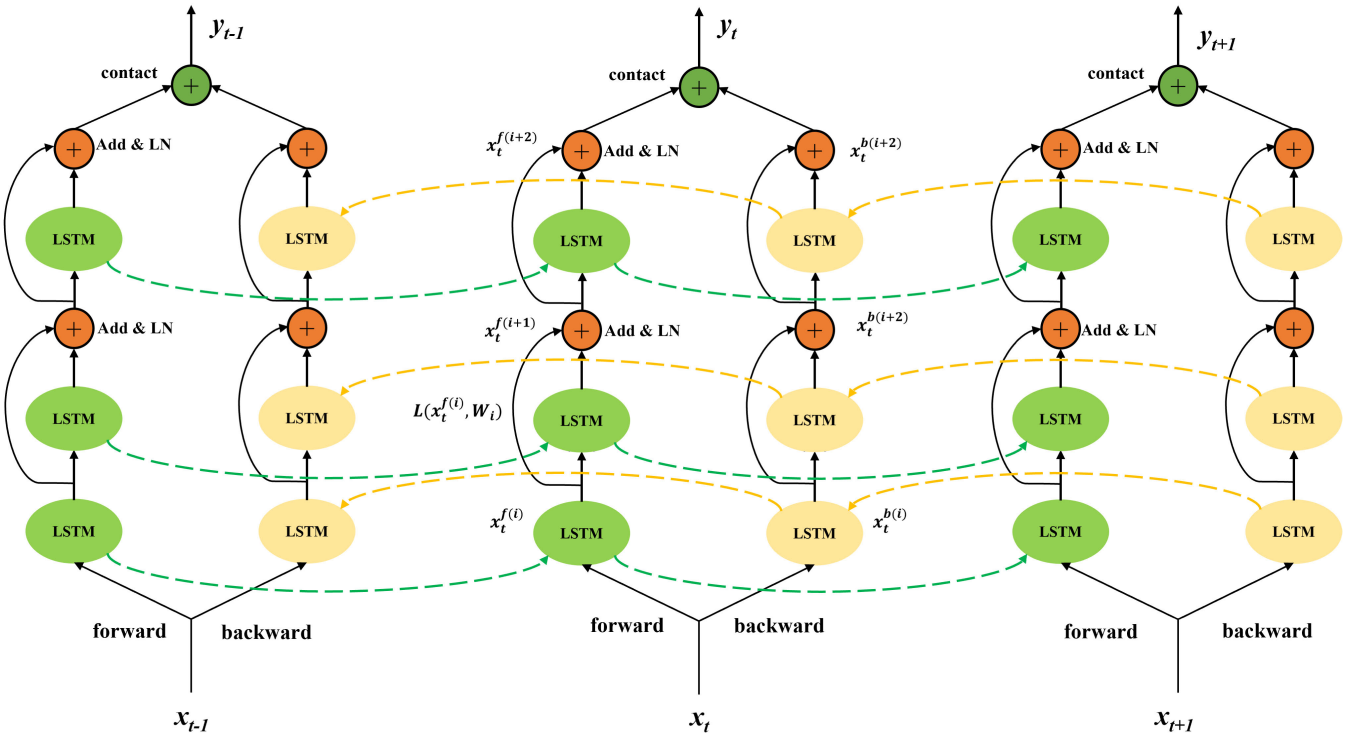


FIGURE 5. The ResBLSTM network is composed of forward and backward LSTM networks, each LSTM is added to the residual structure and LN, and the final encoding information forward state and backward state are spliced.

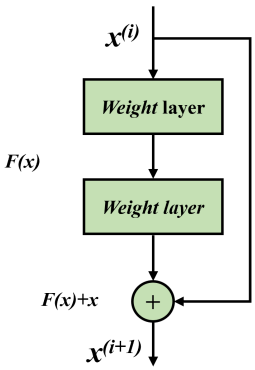


FIGURE 6. The residual block in ResNet.

Hochreiter et al. [39] proposed a long short-term memory network (LSTM). Unlike simple RNNs, LSTM is a gated-based recurrent neural network that can effectively retain longer-term temporal information. Moreover, it outperforms simple RNNs in handling longer time series. Nevertheless, behavioral data is influenced not only by preceding moments but also by subsequent moments. Bidirectional LSTM (BLSTM) is a two-way LSTM network that considers both forward and backward information. Compared to the LSTM network, BLSTM enhances time series feature extraction by capturing bidirectional dependencies. Therefore, utilizing a BLSTM network to extract time series features from behavioral data is an appropriate approach.

Although the BLSTM network can extract time series features well, it is not strong at capturing spatial features, and with the increase in the number of stacking layers, the problem of gradient disappearance will also occur during training. To solve this problem of gradient disappearance, in 2015, the Microsoft Research team built the residual network ResNet [40]. The network reached 152 layers, and it won the championship in ILSVRC in 2015. The specific residual structure is shown in Fig.6. Each residual block can be expressed as:

$$x^{i+1} = x^{(i)} + F\left(x^i, W_i\right). \tag{6}$$

The residual blocks are divided into two parts, where x^i is a direct mapping, $F\left(x^i, W_i\right)$ is the residual part.

Similarly, the structure mentioned above is also employed in the design of the encoder component in the Transformer model. Our research offers a residual structure based on the BLSTM network, which builds on the benefits of this structure. Normalization techniques can also be used in the BLSTM network. Layer normalization (LN) [20] is particularly advantageous for recurrent neural networks compared to batch normalization (BN), LN is computed similarly to BN and can be expressed as follows:

$$\widehat{x^{(i)}} = \frac{x^{(i)} - E(x^{(i)})}{\sqrt{var(x^{(i)})}}. \tag{7}$$

where $x^{(i)}$ represents the input vector of the i dimension, $\widehat{x^{(i)}}$ represents the output after layer normalization.

In this paper, a new combination that combines residual structure and layer normalization in a BLSTM network is called ResBLSTM, as illustrated in Fig.5. The recursive feature information y can be described as:

$$x_t^{f(i+1)} = LN\left(x_t^{f(i)} + L\left(x_t^{f(i)}, W_i\right)\right), \tag{8}$$

$$x_t^{b(i+1)} = LN\left(x_t^{b(i)} + L\left(x_t^{b(i)}, W_i\right)\right), \tag{9}$$

$$y^t = concat(x_t^f, x_t^b). \tag{10}$$

where LN is layer normalization, L is the processing of input states in the LSTM network, the subscript t in $x_t^{f(i+1)}$ represents the t-th moment in the time series, the f in the superscript represents the forward state, b represents the reverse state, and (i+1) represents the number of stacked layers, the encoded information y_t at time t is spliced together from the forward state and the backward state.

C. ATTENTION MECHANISM

The attention mechanism has found broad applications in various domains of deep learning. Models based on the attention mechanism can not only capture the positional relationships among information but also quantify the significance of different information features based on their intrinsic characteristics [41], [42], [43]. When applying RNNs for activity recognition, the typical approaches involve using either the average of feature information $C = \frac{1}{T}\sum_{t=1}^{T} y_t$ or the final feature information $C = y_T$ for identification. These two methods fail to consider the varying importance of different feature information in behavior recognition, highlighting a significant drawback. Therefore, we leverage the attention mechanism introduced by Raffel and Ellis [44] to enhance the accuracy of behavior prediction. As shown in Fig.7, the calculation method for feature information C is as follows:

$$e_t = f(y_t), \quad \alpha_t = \frac{exp(e_t)}{\sum_{i=1}^{T} exp(e_i)}, \quad C = \sum_{t=1}^{T} \alpha_t y_t. \tag{11}$$

where the function f expression can be learned by backpropagation through the fully connected layer, and the weight α_t constantly changing during the training process, and as the training progresses, the feature information C gradually becomes more representational.

V. EXPERIMENTS AND EVALUATIONS

A. EXPERIMENTAL MODELS AND ENVIRONMENT SETTING

In this experiment, the model training framework used is TensorFlow 2.1. The testing was conducted on a computer equipped with an i7 CPU, 16GB of RAM, and an NVIDIA RTX3050 GPU. Table 1 summarizes the relevant parameters of the 1DCNN-ResBLSTM-Attention model.

Our model is trained using a fully supervised approach, where various types of samples are labeled. To measure

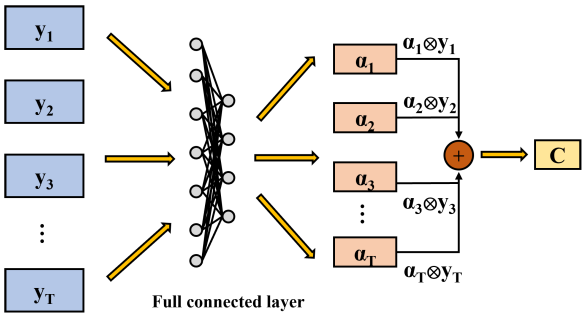


FIGURE 7. The attention mechanism calculates the weight of the recursive information y generated by ResBLSTM, and the calculated weight is α, and each weight α multiplies and sums with y to obtain the final feature information C.

the performance, we utilize the cross-entropy loss between predicted samples and actual samples. The Nadam optimizer is chosen to fit the model, aiming to achieve the best performance. For training, we set the number of epochs to 150 and the learning rate to 0.001. Due to memory constraints on the computer, we set the batch size to 64. In the convolution process, we avoid changing the feature series' dimension through convolution. Instead, we use pooling kernels to reduce the time series' length. The P_Stride, which is the step size of the pooled kernel, is set to 2 for computational convenience. Each time the time series is pooled, its length is halved. To prevent overfitting, dropout is incorporated in the model. Other hyperparameters are kept relatively unchanged to avoid consuming excessive computing resources through excessive fine-tuning. The focus is on maintaining a balance between model performance and computational efficiency.

B. EVALUATION INDEXES

For binary classification problems, evaluation can be done with accuracy, precision, recall, and the F1-score. Accuracy refers to the proportion of all samples that are correctly classified, and higher accuracy means a better classification effect. To measure the overall classification ability of the system for all categories. Precision represents the proportion of samples that truly belong to a category among all samples identified as a category and measures the system's classification accuracy for a category. The recall rate represents the proportion of all samples of a class that are correctly identified as that category and measures the system's comprehensiveness of classification for a class. The F1-score is a harmonic mean of precision and recall,that measures the average classification performance of the system for a category. They are expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{12}$$

$$Precision = \frac{TP}{TP + FP}. \tag{13}$$

$$Recall = \frac{TP}{TP + FN}. \tag{14}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}. \tag{15}$$

TABLE 1. 1DCNN-ResBLSTM-Attention model parameters.

Stage	Hyperparameters	Values used	
Structure	Convolution_1+Maxpooling Dropout rate = 0.2	Kernel-size	3
		K_Stride	1
		Filters	256
		Pool-size	2
		P_Stride	2
	Convolution_2+Maxpooling Dropout rate = 0.2	Kernel-size	3
		K_Stride	1
		Filters	128
		Pool-size	2
		P_Stride	2
	Convolution_3+Maxpooling Dropout rate = 0.2	Kernel-size	3
		K_Stride	1
		Filters	64
		Pool-size	2
		P_Stride	2
	Convolution_4+Maxpooling Dropout rate = 0.2	Kernel-size	3
		K_Stride	1
		Filters	32
		Pool-size	2
		P_Stride	2
	ResBLSTM_1	neuron	64
	ResBLSTM_2	neuron	64
	Attention layer		
	Dense	neuron	6,6,18
Training	Optimizer	Nadam	
	Maximum epochs	150	
	Batch size	64	
	Learning rate	0.001	
	Loss Function	Cross entropy	

where TP means the number of true positives, TN means the number of true negatives, FP means the number of false positives, and FN means the number of false negatives.

For multi-category problems, it is necessary to use macro-averaging, which involves calculating the accuracy, recall, and F1-score for each category and then calculating the arithmetic mean. The macro-averages of accuracy, recall, and the F1-score are calculated as follows:

$$M\_Precision = \frac{1}{N}\sum_{i=1}^{N} Precision_i. \tag{16}$$

$$M\_Recall = \frac{1}{N}\sum_{i=1}^{N} Recall_i. \tag{17}$$

$$M\_F1-score = \frac{1}{N}\sum_{i=1}^{N} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}. \tag{18}$$

where N is the total number of categories, i denotes the i-th category.

C. EXPERIMENTAL RESULTS AND ANALYSES

1) COMPARISON OF RESBLSTM TO BLSTM

We constructed ResBLSTM by incorporating the residual structure and layer normalization techniques from ResNet and Transformer. We experimented with different numbers of stacked layers (3, 5, and 7) for both BLSTM and ResBLSTM on the UCI-HAR dataset. A comparison is made on the UCI-HAR dataset, and the results are shown in Table 2. Similar to convolutional networks, increasing the number of stacked layers enhances the feature extraction capability. For both BLSTM and ResBLSTM, as the number of layers increased, the recognition accuracy improved. Specifically, the recognition accuracy of BLSTM increased from 91.99% to 92.63% when the number of layers went from 3 to 7, while ResBLSTM improved from 95.86% to 96.36%. This demonstrates the benefit of layer stacking for feature extraction. Furthermore, compared to BLSTM, ResBLSTM achieved significantly higher recognition accuracy, with an increase of 3.87% for 3 layers, 3.8% for 5 layers, and 3.73% for 7 layers, along with notable improvements in other metrics. Based on these experimental results, the addition of residual structure and layer normalization to the BLSTM network proves to be highly effective. Therefore, we adopt the ResBLSTM network as the time series feature extraction network in our overall network design, and we anticipate its applicability in various fields.

TABLE 2. Comparison of ResBLSTM networks and BLSTM networks at Layer 3, Layer 5, and Layer 7.

	Accuracy (%)	M_Precision (%)	M_Recall(%)	M_F1-score(%)
BLSTM-3	91.99	92.17	92.12	92.09
ResBLSTM-3	95.86	95.85	95.93	95.85
Increase	↑ 3.87	↑ 3.68	↑ 3.81	↑ 3.76
BLSTM-5	92.33	92.40	92.40	92.39
ResBLSTM-5	96.13	96.31	96.10	96.15
Increase	↑ 3.80	↑ 3.91	↑ 3.70	↑ 3.76
BLSTM-7	92.63	92.67	92.71	92.66
ResBLSTM-7	96.36	96.40	96.39	96.37
Increse	↑ 3.73	↑ 3.73	↑ 3.68	↑ 3.71

2) ABLATION STUDY

This study aims to evaluate the effectiveness of each section using accuracy, precision, recall, and F1-score as the evaluation criteria. Table 3 provides an overview of the results obtained from testing the three datasets using the 1DCNN and ResBLSTM methods separately. The results show that when the time series length is 128, the recognition accuracy of 1DCNN is significantly higher than that of ResBLSTM in all three datasets. However, when the time series length is 300, the recognition accuracy of both methods becomes similar. This suggests that for shorter time series, spatial features such as numerical values play a more important role in behavior recognition, while as the time series length increases, the significance of temporal features becomes more prominent. Therefore, combining 1DCNN with ResBLSTM allows for capturing both spatial and temporal features. When comparing the performance of 1DCNN+ResBLSTM with ResBLSTM alone, the accuracy rates on the UCI-HAR, WISDM, and KU-HAR datasets improve by 1.46%, 1.55%, and 0.61%, respectively. Similarly, when compared to 1DCNN alone, the accuracy rates improve by 0.25%, 0.3%, and 0.95%, respectively.

TABLE 3. Ablation experiments on three publicly available datasets to verify the effectiveness of each plate.

Dataset	Algorithm	Accuracy (%)	M_Precision (%)	M_Recall (%)	M_F1-score (%)	recognition time (ms)
UCI-HAR Time series length (128) Number of sensor channels:(9)	ResBLSTM	95.58	95.82	95.47	95.58	15
	1DCNN	96.79	97.07	96.97	96.99	4
	1DCNN+ResBLSTM	97.04	97.08	97.05	97.06	5
	1DCNN+ResBLSTM+Attention	**98.37**	**98.42**	**98.43**	**98.42**	**6**
WISDM Time series length (128) Number of sensor channels:(3)	ResBLSTM	97.08	96.03	96.32	96.17	15
	1DCNN	98.36	97.56	97.98	97.76	4
	1DCNN+ResBLSTM	98.63	97.98	98.17	98.08	5
	1DCNN+ResBLSTM+Attention	**99.01**	**98.73**	**98.73**	**98.73**	**6**
KU-HAR Time series length (300) Number of sensor channels:(6)	ResBLSTM	96.92	97.40	96.78	97.07	29
	1DCNN	96.58	97.70	97.06	97.33	4
	1DCNN+ResBLSTM	97.53	97.70	97.61	97.65	6
	1DCNN+ResBLSTM+Attention	**97.89**	**98.33**	**98.01**	**98.16**	**7**

Table 3 indicates that the addition of the attention mechanism results in an increase in recognition accuracy of 1.33%, 0.38%, and 0.97% on the UCI-HAR, WISDM, and KU-HAR datasets, respectively. Similar improvements are observed in precision, recall, and F1-score. Overall, these findings highlight the effectiveness of combining 1DCNN and ResBLSTM;at the same time, combined with the attention mechanism, the accuracy and performance of behavior recognition on the test data set can be enhanced.

In addition to accuracy, efficiency is also an important factor to consider. In this study, the recognition time of individual actions is used as a criterion for efficiency evaluation. Utilizing the parallel operation capability of the GPU, the one-dimensional convolution operation for a single behavior takes only 4ms. On the other hand, ResBLSTM is a recurrent neural network, and its calculation time is dependent on the length of the time series. With a time series length of 128, the recognition time for a single action using ResBLSTM is 15ms. When the time series length is 300, the recognition time increases to 29ms. This highlights the drawback of ResBLSTM's relatively long recognition times. To address the issue of lengthy recognition times in ResBLSTM, the combination of 1DCNN and ResBLSTM, denoted as 1DCNN+ResBLSTM, provides a solution. When applied to a time series length of 128, the recognition time for a single action is reduced from 15ms to 5ms. Similarly, when applied to a time series length of 300, the recognition time decreases from 29ms to 6ms. It is worth noting that the addition of an attention mechanism slightly increases the recognition time. However, the increase in accuracy compensates for this small trade-off, making it an acceptable trade-off. The combination of 1DCNN and ResBLSTM, along with the incorporation of an attention mechanism, not only improves accuracy but also enhances efficiency by significantly reducing the recognition time of individual actions on both short and long-time series.

3) COMPARE EACH BEHAVIOR

Distinguishing similar actions poses a significant challenge in action recognition, as the data patterns of these actions are often similar and difficult to differentiate. To validate the effectiveness of the proposed model, we conducted a thorough analysis of its performance on each action. To evaluate the model's performance on each action, we utilize the F1-score as a balanced evaluation metric. The F1-score takes into account both precision and recall, providing a comprehensive measure of performance that considers false positives and false negatives. The WISDM dataset is to put the mobile phone in the trouser pocket and use the mobile phone's triaxial accelerometer to collect data.In this dataset, there are three sets of actions that are highly similar: upstairs and downstairs, standing and sitting, and walking and jogging. Since the mobile phone is placed in the pants pocket, the leg movements during jogging exhibit more obvious frequency and amplitude compared to walking. Similarly, for the actions of standing and sitting, the amplitude of leg movement is reversed. Both ResBLSTM and 1DCNN achieve recognition accuracies of over 98% for these two sets of movements. However, in the case of going upstairs and going downstairs, the leg movements exhibit smaller changes in amplitude and frequency. The recognition accuracy of ResBLSTM for this set of similar movements is less than 90%(showed in Fig.8(a)). Conversely, 1DCNN outperforms ResBLSTM significantly in recognizing these movements, indicating that 1DCNN can capture subtle changes in data amplitude more effectively. The combination of 1DCNN with ResBLSTM shows a slight improvement in the performance of each action, highlighting the importance of considering both spatial and temporal characteristics in action recognition. The results suggest that the joint utilization of spatial and temporal features can enhance the model's ability to distinguish similar actions accurately.

The UCI-HAR dataset involves placing the mobile phone on the waist and collecting data using the accelerometer and gyroscope. For the actions of standing and sitting, the changes in waist amplitude are smaller compared to leg movements. As illustrated in Fig.8(b), the recognition accuracy of a single feature extraction model, whether it is 1DCNN or ResBLSTM, is significantly reduced for these actions. However, in comparison to the WISDM dataset, the

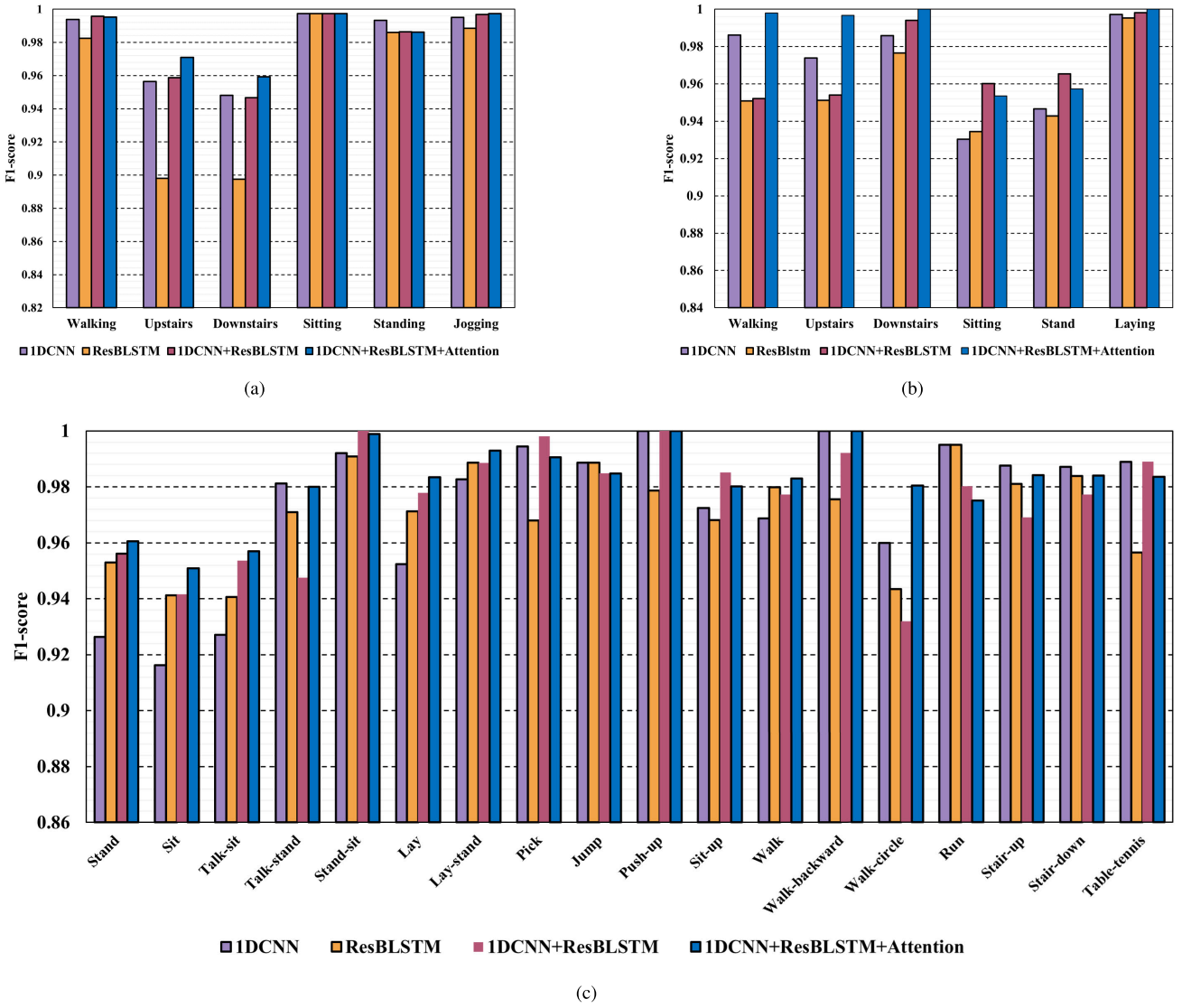


FIGURE 8. (a) , (b), (c) are the F1-scores for each action of the four models in the WISDM, UCI-HAR, KU-HAR data, respectively.

recognition accuracy for upstairs and downstairs movements improves in the UCI-HAR dataset. This improvement can be attributed to the important role played by the angle changes captured by the gyroscope in these particular actions. For the actions of standing and sitting, the combination of 1DCNN and ResBLSTM proves effective. After combining with ResBLSTM, the F1-score increases by 2.24% and 2.58%, respectively, compared to using 1DCNN alone. However, in the case of walking and going upstairs, the performance of the combined model is not as good as that of 1DCNN alone. The addition of the attention mechanism leads to a significant improvement in recognition accuracy. When compared to 1DCNN+ResBLSTM, the F1-score increases by 4.58% and 4.28%, respectively. This indicates that the attention mechanism is highly effective in calculating the weights of feature information.

The KU-HAR and UCI-HAR datasets have similar data acquisition methods, with a time series length of 300 for a single behavior. However, these datasets differ from WISDM and UCI-HAR in that they contain more similar and transitional actions. Actions such as stand, sit, and stand-sit share a lot of similarities (shown in Fig.8(c)). In these actions, ResBLSTM demonstrates higher recognition accuracy compared to 1DCNN, highlighting the advantages of temporal features in capturing action sequences of length 300. The combined recognition accuracy of 1DCNN and ResBLSTM improves for these actions, although the combination exhibits some instability in certain actions like talk-stand and walk-circle. However, the addition of the attention mechanism enhances the stability of the overall model. For 10 out of the 18 actions, the highest recognition accuracy is achieved, and the remaining 8 actions have an F1-

TABLE 4. Classification confusion matrix on WISDM.

Activity		Predict Label					
		Walk	Upstairs	Downstairs	Sit	Stand	Jog
True Label	Walk	1277	1	5	0	0	1
	Upstairs	0	406	6	0	0	2
	Downstairs	2	10	293	0	0	0
	Sit	0	0	0	186	0	0
	Stand	1	0	0	0	145	0
	Jog	4	1	1	0	0	1091

TABLE 5. Classification confusion matrix on UCI-HAR.

Activity		Predict Label					
		Walk	Upstairs	Downstairs	Sit	Stand	Lay
True Label	Walk	496	0	0	0	0	0
	Upstairs	2	469	0	0	0	0
	Downstairs	0	0	420	0	0	0
	Sit	0	1	0	472	18	0
	Stand	0	0	0	27	505	0
	Lay	0	0	0	0	0	537

TABLE 6. Classification confusion matrix on KU-HAR.

Activity		Predict Label																	
		Std	Sit	Talk-sit	Talk-std	Std-sit	Lay	Lay-std	Pick	Jump	Push-up	Sit-up	Walk	Walk-back	Walk-circle	Run	Up	Down	Table-tennis
True Label	Std	365	11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	Sit	4	368	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	Talk-sit	10	13	334	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
	Talk-std	2	0	1	368	0	0	0	0	0	0	0	2	0	0	0	0	0	0
	Std-sit	1	0	0	0	435	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lay	0	6	2	0	0	355	0	0	0	0	0	0	0	0	0	0	0	0
	Lay-std	0	0	0	3	0	1	348	0	0	0	0	0	0	0	0	0	0	0
	Pick	0	0	0	0	0	0	0	263	0	0	4	0	0	0	0	0	0	0
	Jump	0	0	0	0	0	0	0	1	130	0	0	0	0	0	1	0	1	0
	Push-up	0	0	0	0	0	0	0	0	0	96	0	0	0	0	0	0	0	0
	Sit-up	1	1	0	0	0	0	1	0	0	0	198	0	0	0	0	0	0	0
	Walk	0	0	0	3	0	0	0	0	0	0	0	173	0	0	0	0	0	0
	Walk-back	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0	0	0
	Walk-circle	0	0	0	1	0	0	0	0	0	0	0	1	0	50	0	0	0	0
	Run	0	0	0	0	0	0	0	0	1	0	0	0	0	0	98	0	0	0
	Up	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	156	2	0
	Down	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	154	0
	Table-tennis	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	90

TABLE 7. Comparison between the proposed algorithm and the existing algorithm in UCI-HAR, WISDM, and KU-HAR dataset.

Algorithm	UC-HAR		WISDM		Algorithm	KU-HAR	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)		Accuracy (%)	F1-score (%)
LSTM [45]	95.38	-	97.52	-	DeepTransHHAR [46]	-	94.25
LSTM-CNN [32]	-	95.78	-	95.85			
FMR [47]	89.48	89.81	96.83	95.79	DeepConv [48]	96.67	96.41
CNN-BiLSTM [49]	96.31	96.37	96.05	96.04			
CNN-GRU [50]	96.20	96.19	97.21	97.22	ResNet [24]	93.54	-
Transformer [28]	95.38	95.37	98.18	97.20			
Our proposed	**98.37**	**98.42**	**99.01**	**98.37**	**Our proposed**	**97.89**	**98.16**

score of over 98%. These results indicate that the combination of 1DCNN, ResBLSTM, and Attention is highly effective in distinguishing similar and transitional actions. It improves the stability of the model and enables accurate recognition of actions that are otherwise challenging to differentiate.

4) CONFUSION MATRIX ON PUBLIC DATASETS

The confusion matrix provides insight into the recognition results of each action, where the horizontal axis represents the predicted results and the vertical axis represents the true labels. Table 4 shows the confusion matrix of the WISDM

test set, consisting of 3432 data samples. The dataset has a relatively larger proportion of walking and jogging actions. Among the 1284 walking actions, only 7 were misclassified, and among the 1097 jogging actions, 6 were misclassified, resulting in recognition accuracies exceeding 99.4%. The recognition accuracy for standing and sitting actions reached 99.9%. As analyzed in the previous section (compare each behavior), the prediction errors were primarily concentrated in the upstairs and downstairs actions. Specifically, 10 downstairs actions were misclassified as upstairs, and 6 upstairs actions were misclassified as downstairs. Despite these errors, the overall recognition accuracy still reached 99.01%.

Table 5 represents the confusion matrix of the UCI-HAR test set, consisting of more than 400 samples for each action. The data distribution in this dataset is relatively uniform. Out of the total 2947 action data samples, 2926 were correctly classified, resulting in an accuracy rate of 98.38%. Unlike the WISDM dataset, the majority of misclassifications occur in the standing and sitting actions, with recognition accuracies of 96.13% and 94.92%, respectively. A significant portion of these recognition errors can be attributed to the mutual misclassification between these two actions. The reason behind these errors is that when the smartphone is placed on the waist, the difference between the standing and sitting actions becomes less pronounced, leading to higher confusion between the two actions.

Table 6 displays the confusion matrix of the KU-HAR dataset, which consists of a richer variety of action types. Despite the increased complexity, the combination of 1DCNN+ResBLSTM+Attention is still able to effectively identify actions with higher discriminability. The recognition errors are mainly concentrated on three actions: stand, sit, and talk-stand. Among these three actions, a total of 40 mutual recognition errors occur. It is worth noting that for these similar actions, the performance of 1DCNN+ResBLSTM+Attention is relatively poor compared to other actions. However, even in these challenging scenarios, the model still achieves a recognition accuracy of more than 93%.

D. COMPARISON OF PROPOSED ALGORITHM WITH PREVIOUS

To demonstrate the superior performance of our proposed model, we conducted a comparative analysis with existing algorithms that have been tested on the WISDM and UCI-HAR datasets in recent years. In Table 7, we present the comparison results using the same evaluation metrics to ensure the reliability and fairness of the comparison. From Table 7, it is evident that our proposed model outperforms the existing algorithms on both the WISDM and UCI-HAR datasets. The performance metrics such as accuracyand F1-score demonstrate that our model has achieved the best results among the evaluated algorithms. These findings validate the effectiveness and superiority of our proposed model for action recognition tasks on these public datasets.

The KU-HAR dataset, which was constructed in 2021, has a limited number of existing methods available for comparison. In Table 7, we compared the performance of our proposed model with a few other approaches that have been tested on the KU-HAR dataset. The evaluation metrics utilized for comparison demonstrate that our model exhibits superior accuracyand F1-score compared to the other methods. This indicates that our proposed model is highly effective in accurately recognizing actions on the KU-HAR dataset.

VI. CONCLUSION

In this paper, we propose a human activity recognition model that combines three key components: a one-dimensional convolutional neural network (1DCNN), an improved bidirectional long short-term memory network (ResBLSTM), and attention mechanisms. The proposed model aims to improve the accuracy and stability of activity recognition by effectively capturing both spatial and temporal features of the input data. To begin with, the model utilized a one-dimensional convolutional neural network to extract the spatial features of the input data. The 1DCNN applies convolutional operations to capture patterns and structures in the data, and a pooling layer with an appropriate step size is employed to reduce the length of the time series while preserving relevant information. Next, the improved ResBLSTM network is employed to extract temporal features from the data. The bidirectional nature of the ResBLSTM allows it to effectively model the dependencies and dynamics of the sequential data. By considering information from both past and future time steps, the ResBLSTM captures important temporal patterns and context in the activity sequences. Finally, the attention mechanism is incorporated into the model to calculate the weights of different feature information in the final recognition process. The attention mechanism assigns higher weights to more informative features, thereby enhancing the discriminative power of the model and improving the accuracy and stability of activity recognition. By combining the strengths of the 1DCNN, ResBLSTM, and attention mechanisms, our proposed model aims to overcome the limitations of existing approaches and achieve improved performance in human activity recognition tasks.

The ResBLSTM network, proposed in this study, exhibits superior feature extraction capability and training stability when compared to the BLSTM network. By combining the 1DCNN and ResBLSTM models, we enhance the feature extraction methods, enabling better identification of similar human activities. Furthermore, the incorporation of an attention mechanism enhances the representation ability of the final recognition information, further improving the model's ability to distinguish between different actions.

In the UCI-HAR public dataset, we constructed ResBLSTM and BLSTM networks with different stacking layers. The ResBLSTM network achieved a significant improvement of approximately 3.6% in the F1-score compared

to the BLSTM network. Additionally, we evaluated the performance of our proposed 1DCNN-ResBLSTM-Attention model on three public datasets: UCI-HAR, WISDM, and KU-HAR. The overall recognition accuracy obtained was 98.37%, 99.01%, and 97.89%, respectively. Ablation experiments were conducted to assess the effectiveness of each component in our proposed model. These experiments compared the F1-scores of different actions across the three datasets. The results clearly demonstrated that our proposed model excelled at distinguishing similar actions, providing evidence of its efficacy.

The research has two primary limitations: (1) Insufficient discussion about the optimal combination of convolutional and Res-BLSTM layers for different datasets, and the importance of fine-tuning other hyperparameters for achieving better results. (2) Higher algorithmic complexity compared to LWRF and local binary methods, leading to increased computing resource consumption in HAR. Additionally, the method's recognition performance may decrease when test data contains singular values.

In this study, all experiments were conducted using human activity datasets collected from mobile phones. As part of future research directions, we plan to extend our experiments to include different wearable sensing devices. Additionally, we aim to explore methods to reduce the computational requirements of the model and minimize hardware power consumption.

REFERENCES

[1] M. Sjöström, U. Ekelund, A. Yngve, M. J. Gibney, B. Margetts, J. M. Kearney, and L. Arab, "Assessment of physical activity in youth," *J. Appl. Physiol.*, vol. 105, no. 3, pp. 977–987, 2004.

[2] W. S. Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama, "Human activity recognition using inertial sensors in a smartphone: An overview," *Sensors*, vol. 19, no. 14, p. 3213, Jul. 2019.

[3] Y.-W. Kim, K.-L. Joa, H.-Y. Jeong, and S. Lee, "Wearable IMU-based human activity recognition algorithm for clinical balance assessment using 1D-CNN and GRU ensemble model," *Sensors*, vol. 21, no. 22, p. 7628, Nov. 2021.

[4] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2004, pp. 719–722.

[5] V. Michalis, N. Christophoros, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers Robot. AI*, vol. 2, no. 28, p. 28, 2015.

[6] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A survey on activity detection and classification using wearable sensors," *IEEE Sensors J.*, vol. 17, no. 2, pp. 386–403, Jan. 2017.

[7] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–40, 2022.

[8] Z. Feng, L. Mo, and M. Li, "A random forest-based ensemble method for activity recognition," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 5074–5077.

[9] A. Jain and V. Kanhangad, "Human activity classification in smartphones using accelerometer and gyroscope sensors," *IEEE Sensors J.*, vol. 18, no. 3, pp. 1169–1177, Feb. 2018.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Services*, Nov. 2014, pp. 197–205.

[12] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[13] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[14] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, Nov. 2017.

[15] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. Esann*, vol. 3, Apr. 2013, p. 3.

[16] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[18] D. Buffelli and F. Vandin, "Attention-based deep learning framework for human activity recognition with user adaptation," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13474–13483, Jun. 2021.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[20] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[21] T. Asuroglu, K. Açici, Ç. B. Erdas, and H. Ogul, "Texture of activities: Exploiting local binary patterns for accelerometer data analysis," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2016, pp. 135–138.

[22] T. Asuroglu, "Complex human activity recognition using a local weighted approach," *IEEE Access*, vol. 10, pp. 101207–101219, 2022.

[23] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Exp. Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.

[24] S. Mekruksavanich, P. Jantawong, N. Hnoohom, and A. Jitpattanakul, "ResNet-based network for recognizing daily and transitional activities based on smartphone sensors," in *Proc. 3rd Int. Conf. Big Data Analytics Practices (IBDAP)*, Sep. 2022, pp. 27–30.

[25] S. Ishimaru, K. Hoshika, K. Kunze, K. Kise, and A. Dengel, "Towards reading trackers in the wild: Detecting reading activities by EOG glasses and deep neural networks," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 704–711.

[26] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 351–360.

[27] Y. Shavit and I. Klein, "Boosting inertial-based human activity recognition with transformers," *IEEE Access*, vol. 9, pp. 53540–53547, 2021.

[28] Z. N. Khan and J. Ahmad, "Attention induced multi-head convolutional neural network for human activity recognition," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107671.

[29] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[30] E. Büber and A. M. Guvensan, "Discriminative time-domain features for activity recognition on a mobile phone," in *Proc. IEEE 9th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process. (ISSNIP)*, Apr. 2014, pp. 1–6.

[31] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, Apr. 2014.

[32] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[33] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

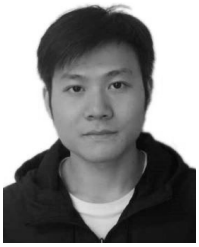
[34] N. Sikder and A.-A. Nahid, "KU-HAR: An open dataset for heterogeneous human activity recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 46–54, Jun. 2021.

[35] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*.

[36] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.
[37] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[38] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Planning Inference*, vol. 90, no. 2, pp. 227–244, Oct. 2000.
[39] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[42] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[44] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," 2015, *arXiv:1512.08756*.
[45] M. Zhang, "Gait activity authentication using LSTM neural networks with smartphone sensors," in *Proc. 15th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, Dec. 2019, pp. 456–461.
[46] P. Kumar and S. Suresh, "DeepTransHHAR: Inter-subjects heterogeneous activity recognition approach in the non-identical environment using wearable sensors," *Nat. Acad. Sci. Lett.*, vol. 45, no. 4, pp. 317–323, Aug. 2022.
[47] H. Xu, J. Li, H. Yuan, Q. Liu, S. Fan, T. Li, and X. Sun, "Human activity recognition based on gramian angular field and deep convolutional neural network," *IEEE Access*, vol. 8, pp. 199393–199405, 2020.
[48] N. Sikder, M. A. R. Ahad, and A.-A. Nahid, "Human action recognition based on a sequential deep learning model," in *Proc. Joint 10th Int. Conf. Informat., Electron. Vis. (ICIEV), 5th Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, Aug. 2021, pp. 1–7.
[49] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *Vis. Comput.*, vol. 38, no. 12, pp. 4095–4109, Dec. 2022.
[50] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input CNN-GRU based human activity recognition using wearable sensors," *Computing*, vol. 103, no. 7, pp. 1461–1478, Jul. 2021.

JUNJIE ZHANG was born in Wuhan, China, in April 1980. He received the Ph.D. degree in automation from the University of Lille, France, in 2017.

He is a Professor and the Master's Supervisor with the School of Computer and Artificial Intelligence, Wuhan Textile University. He specializes in researching the clothing digitization and artificial intelligence. With a garment recommendation and artificial intelligence in the field, he has authored over 30 articles and secured four patents, solidifying his contributions to the field.



YUANHAO LIU received the B.E. degree in electrical engineering and automation from the Hubei University of Automotive Technology, Shiyan, China, in 2020. He is currently pursuing the M.S. degree with the School of Computer and Artificial Intelligence, Wuhan Textile University, Wuhan, China. His research interests include machine learning, artificial intelligence, activity recognition, and image restoration.

HUA YUAN was born in Wuhan, China, in July 1982. She received the Ph.D. degree in economics from the Wuhan University of Technology, Wuhan, in 2019.

Following the completion of the Ph.D. study, she began her career as a Lecturer with the Fashion School, Wuhan Textile University. She specializes in researching the sustainable development of the fashion industry, behavior of fashion consumption, and productivity within the fashion sector. She has a commitment to advancing knowledge and fostering innovation in the field, she has authored over 20 articles and secured two patents, solidifying her contributions to the field.