Attention-Based Multihead Deep Learning Framework for Online Activity Monitoring With Smartwatch Sensors

Dipanwita Thakur, *Member, IEEE*, Antonella Guzzo, *Senior Member, IEEE*, and Giancarlo Fortino, *Fellow, IEEE*

***Abstract*—The expeditious propagation of Internet of Things (IoT) technologies implanted in different smart devices, such as smartphones and smartwatches have a ubiquitous consequence on the modern population. These devices are employed to collect data and to aid in tracking and analyzing the users' daily activities using various human activity monitoring and recognition (HAMR) techniques. However, most current HAMR approaches rely on exploratory case-based shallow feature learning architectures, which endeavor to recognize activities correctly in real-world situations. To address this issue, we offer a unique strategy for HAMR that leverages the attention mechanism with multihead convolutional neural networks (CNNs) and long-short-term-memory (LSTM). The accuracy of activity detection is improved in the presented method by integrating attention into multihead CNNs followed by LSTM for better feature extraction and selection. Verification investigations are carried out using data from the University of California (UCI) repository, which is publicly available. The results show that our proposed framework is more accurate than current frameworks using both the 10-fold and leave-one-subject-out cross-validation. Finally, the proposed method can recognize human activity in real time, regardless of the type of smart device.**

***Index Terms*—Attention mechanism, deep learning, human activity monitoring and recognition, Internet of Healthcare Things (IoHT).**

I. Introduction

THE FAST growth of wearable and mobile technologies has made it simple to apply Internet of Things (IoT) technology in healthcare. Real-time human activity monitoring, especially for older people's activities of daily living (ADL), is a crucial problem in smart healthcare. Mobile and wearable sensors might enhance senior care and medical rehabilitation. In order to better understand people's daily behavior and interactions with their living environments, human activity monitoring and recognition (HAMR) in pervasive computing environments has become a hotly debated topic. This topic has been thoroughly investigated for the so-called Internet of Healthcare Things (IoHT) [1].

Including HAMR, inertial sensors can be used to better understand human behavior, which could open up many doors for IoHT applications in various systems (such as medical healthcare, smart home, smart city, smart transportation, and smart manufacturing). For instance, successful monitoring of the young and the elderly in smart homes necessitates a thorough understanding of residents' everyday home-life activity, which might be used to control smart appliances effectively [1]. Smart devices, such as smartwatches and smartphones that consist of built-in inertial sensors (accelerometer and gyroscope) are used to collect and transmit real-time human activity data through wireless sensor networks due to their low cost and nonintrusive human feature [2].

Earlier, most traditional HAMR techniques employ extensive feature engineering to capture the vibrant signal representations, which are then fed into machine learning for categorization [3]. However, such techniques depend on expert domain knowledge, and the typical use of shallow features reduces the computation efficiency of IoHT-based HAMR. Recently, various deep learning models demonstrated remarkable performance for a range of HAMR tasks by learning high-level feature representation directly from the raw data of smart device sensors, which improves HAMR performance in comparison to conventional methods [4]. Even though current DL studies for HAMR have had much success in many IoHT-related applications, the generalization capacity of the HAR models, or the effectiveness of applying the trained models to a new, unseen, untested data set, remains a crucial difficulty. The variability of the mobile devices disturbances, such as movement artifacts, baseline noise, the occurrence of new activities, different hardware configurations, and user differences, has an impact on the identification rate of the HAMR system and prevents the generalization capability. The model's performance is likely to decline when tested on a new end user whose activity data are never seen in the training set. Moreover, if the unseen, independent data set is from any other smart device and test online then the performance of the model is likely to drop. As a result, a compelling need is to develop a novel automated technique that can efficiently extract essential features and identify and classify human behaviors in IoHT contexts. This article proposes attention-based multihead

Manuscript received 21 February 2023; revised 28 April 2023; accepted 15 May 2023. Date of publication 18 May 2023; date of current version 9 October 2023. This work was supported in part by the WP9.6: "KRR Frameworks for Green-Aware AI" through the PNRR Project FAIR—Future AI Research under Grant PE00000013 and Spoke 9—Green-Aware AI through the NRRP MUR Program funded by the NextGenerationEU, and in part by The PRIN COMMON-WEARS Project funded by the Italian Minister of Research under Grant CUP:H23C22000230005. *(Corresponding author: Antonella Guzzo.)*

Dipanwita Thakur is with the Department of Computer Science, Banasthali Vidyapith, Jaipur 304022, India (e-mail: tdipanwita@banasthali.ac.in).

Antonella Guzzo and Giancarlo Fortino are with the Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, 87036 Rende, Italy (e-mail: antonella.guzzo@unical.it; giancarlo.fortino@unical.it).

Digital Object Identifier 10.1109/JIOT.2023.3277592

convolutional neural network (CNN), followed by long-short-term-memory (LSTM), a new DL model for HAMR that comprises three steps to close that gap. A six-head CNN extracts local features from the tri-axial accelerometer and tri-axial gyroscope in the first step. However, all the extracted local features are not equally important to identify efficiently. Thus, in the second step, an attention mechanism is used to adjust the weights of the features to retain only the most important features from the tri-axial accelerometer and gyroscope from different modalities. Then, using LSTM, the proposed network learns high-level representations and encodes the temporal correlations of the learned local features [5]. LSTM best extends temporal characteristics and processes sequential data, while CNN is best for spatial abstraction and generalization. Therefore, it is crucial to examine the relationship between the human activity's spatial properties and the corresponding information in the time dimension while building a human activity prediction model. The correlation between multidimensional data is mined using CNN, which filters out noise and unstable components. For long sequence prediction, LSTM leverages the information processed by CNN.

In our previous work [6], we proposed an attention-based DL architecture for hemiplegia gait detection, in which the attention mechanism is used with the CNN and LSTM. In this work, we use multihead attention-based CNN with different filter sizes for all the axes of the accelerometer and gyroscope to learn discriminant features from a smartwatch-based data stream for activity recognition.

The primary advantage of the proposed deep learning model over the present HAMR techniques is the significant performance gain achieved by meaningful feature learning and efficient recognition of physical activities. Unlike multihead convolutional attention approaches like those described in [7] and [8], the proposed method does not use LSTM to represent temporal dependency for feature learning which is essential for time-series data. The primary contributions of this article are as follows.

1) The attention-based multihead deep learning framework is proposed for the first and most accurate HAMR in the context of IoHT.
2) A multihead CNN is proposed to learn discriminant features from smartwatch-based tri-accelerometer and tri-gyroscope parallelly with different filter sizes to improve the performance.
3) The multihead convolutional attention mechanism is introduced to adjust the weights of the local features to retain only the essential features.
4) Finally, LSTM is used to represent temporal dependency for feature learning and enhance recognition performance of physical activities.
5) The proposed activity recognition mechanism is implemented offline and online. In offline mode, the smartwatch-based collected data is sent to the PC for activity recognition. In online mode, the activity recognition mechanism is implemented on the smartphone. The classification is done in real time using an unseen, independent data set on an Android smartphone.

II. Related Work

We divide the prior work pertaining to our contributions into three sections: 1) deep learning-based human activity recognition (HAR) (Section II-A); 2) multihead CNN-based HAR (Section II-B); and 3) attention models (Section II-C).

A. Deep Learning for HAR

Recent studies [4] have provided a taxonomy that divides deep learning systems for sensor-based HAR into three different groups. The first group consists of only CNN-based architectures. Only LSTM-based architectures fall into the second category, which is further separated into the LSTM and convolutional LSTM subcategories of models. The raw data was directly fed to the LSTM models proposed in [9], [10], and [11]. In these works, the authors mentioned the limitations of "recurrent neural networks" (RNNs), particularly in long-term dependencies, and used LSTM to overcome those limitations. Though, in HAR literature several approaches proposed using either CNN or LSTM, it has been seen that none of them are perfect for all types of activities, such as static and dynamic. As a result, the researchers started working on different variations of LSTM. Zebin et al. [9] offered a context-aware HAR framework. The authors used body-worn inertial sensors to distinguish between static and dynamic physical activities using a "multilayer LSTM with batch normalization." However, the computational cost and memory needs were rather high in this study since edge computing was employed. With the University of California (UCI)-HAR data set, Yu and Qin [10] proposed a HAR framework based on bidirectional LSTM. Bidirectional LSTM is a slower model and requires more time for training. Zhao et al. [11] proposed a residual bidirectional LSTM architecture to recognize various human behaviors using the UCI smartphone and body-worn sensor (OPPORTUNITY) data sets. However, training LSTM on raw sensory input with a high-sampling rate is not practical due to normal memory and processing resource constraints. They perform poorly because they do not simultaneously leverage temporal and spatial information. Convolutional LSTM models maintain spectral structural locality in their representation. It replaces the inner product of the LSTM with convolutions. Ye et al. [12] proposed two-stream convolutional LSTM and achieved 93.9% accuracy. Two individual spatial-stream and temporal-stream ConvNets are used in this using video data sets to recognize human activities. The third category focuses on hybrid models that use both the CNNs and LSTMs [13], [14], [15], [16]. The correlation between multidimensional data is mined using CNN, which filters out noise and unstable components. For long sequence prediction, LSTM leverages the information processed by CNN. Finally, other deep learning techniques used for HAR are InnoHAR [17], and Multivariate LSTM-FCNA [18]. The authors validated the model using single-device data. Independent, unseen data were not tested yet. Hence, we can say the proposed models suffered from generalization issues.

TABLE I
STATE-OF-THE-ART HAR SYSTEMS

Reference	Device	Sensor	Dataset	Classifier	Accuracy
[19]	Smartphone	Accelerometer Gyroscope	UCI-HAR [20]	multi-head CNN-LSTM	95.76%
[21]	Smartphone	Accelerometer Gyroscope	UCI-HAR [20]	multi-head CNN-Bidirectional LSTM	98%
[7]	Smartphone	Accelerometer	WISDM-HAR [22]	multi-head convolutional attention-based architecture	96.4%
[23]	Smartphone SmartWatch	Accelerometer Gyroscope	UCI-WISDM [24]	Attention-based multi-head CNN-GRU	84.8%(F1-Score)
[8]	Smartphone	Accelerometer Gyroscope	UCI-HAR WISDM-HAR[22]	attention-based multi-head CNN architecture	95.38% 98.18%
[25]	Smartphone	Accelerometer Gyroscope	UCI-HAR [20] WISDM-HAR [22]	CNNbased BiLSTM parallel model with an attention mechanism	96.71% 95.86%
[26]	Smartwatch	Accelerometer Gyroscope	KU-HAR [27]	attention-based Transformer model	99.2%
[28]	Body-worn	Accelerometer Gyroscope	DaLiAc [29]	hybrid attention-based deep neural network	94.55%

B. Multihead CNN-LSTM for HAR

There is very little evidence in HAR literature regarding multihead CNN-LSTM. Ahmad et al. [19] presented a multihead CNN-LSTM architecture to identify six activities and achieved 95.76% accuracy. However, the filter size is fixed for each convolutional layer. Ni et al. [21] proposed a multi-head CNN-LSTM without attention and got an accuracy and F1-score of 98%, using the UCI HAR data set. The used data set is well processed and consists of 561 statistical features. Moreover, in [21], bidirectional LSTM has been used, which is slow to train.

C. Attention-Based Models

Zhang et al. [7] proposed a multihead convolutional attention-based architecture, validated its performance using the WISDM data set, and achieved 96.4% testing accuracy. They have not considered either sensor fusion or different devices to validate their model. Buffelli and Vandin [23] proposed a purely attention-based mechanism for HAR, which replaced RNN. The authors merged the convolutional network with an 8-headed attention layer and used WISDM-UCI [24] data set. In this architecture, layer normalization was used after the attention layer and the fence layer, which enhanced the complexity of the model. Khan and Ahmad [8] proposed attention-based multihead CNN architecture using both the WISDM and UCI data sets. The authors did not mention the required number of epochs to converge. However, they set the number of epochs as 260, which increased the training time of the model. Hence, unsuitable for real-time applications. Yin et al. [25] proposed a CNN-based BiLSTM parallel model with an attention mechanism and achieved 96.71% and 95.86% accuracy using the UCI and WISDM data sets, respectively. However, to achieve these many accuracies the authors adopted a handling model which is almost impossible to use in low-energy efficient and memory-based devices. Luptáková et al. [26] proposed an attention-based Transformer model and achieved an average accuracy of 99.2% using Ku-HAR data where the smartphone's location is fixed. In this work, the authors added normalization layers three times, which may increase the computational complexity of the model. Moreover, the testing accuracy with unseen, independent data is not presented in this work. Zhou et al. [28] proposed a hybrid attention-based deep neural network where feature compression and reconstruction module was used separately.

Many attention-based HAR systems have been explored for healthcare use during the last few decades as shown in Table I. However, all of them mitigate the effect of heterogeneity using any of the preprocessing techniques. Moreover, in HAR, attention models have only been used in addition to a CNN or multihead CNN and not as a means to both the multihead CNN and LSTM, which is the approach we propose in this work.

III. PROPOSED METHODOLOGY

A. Data Collection

The publicly available smartwatch-based raw sensor data is used in this study that is downloaded from the UCI repository [24]. This is a heterogeneous HAR data set. The smartphone and smartwatch-based accelerometer and gyroscope sensors were used to collect the data set. Eight smartphones (two Samsung Galaxy S3 mini, two Samsung Galaxy S3, two LG Nexus 4, and two Samsung Galaxy S+) and four different smartwatches (two LG watches, and two Samsung Galaxy Gears) are used to collect the data from nine users' age range from 25 to 30 years. The users were asked to perform six different activities, such as "Biking," "Sitting," "Standing," "Walking," "Stair Up," and "Stair down." In this study, we use only the smartwatch-based accelerometer and gyroscope data set. Each user conducted five minutes of each activity, which ensured a near-equal data distribution among activity classes (for each user and device). The two different smartwatches were used for diverse sensing scenarios. The smartwatches yielded different maximum sampling frequencies: 200 Hz for LG G and 100 Hz for Samsung Galaxy Gears. Furthermore, the devices exhibit different accelerometer biases and gains.

The data collection was controlled by a custom-made application that ran on the smartwatch. The total number of instances in this smartwatch-based data set is 2738449.

B. Data Preprocessing and Segmentation

To see the impact of the heterogeneity on HAR, no data preprocessing methods are applied in this study. The data set is collected using various smart devices with different sampling rates. In the literature, sub- or super-sampling is often used as a preprocessing technique before training or applying a HAR system to mitigate the heterogeneity effects [24]. In real-time, online HAR, it is not necessary to get the data from a specific device with the same sampling frequency. Using any of the data preprocessing techniques, we mitigate the impact of heterogeneity on HAR. To see the impact of the data heterogeneity on HAR, we investigate the effects without any sub- or super-sampling techniques to train the proposed HAR. In line with the standard approaches [24], we employ a sliding window approach that overcomes the need for explicit semantic segmentations. For processing the time-series data in the HAR problem, a 10-s sliding window with 50% of the overlapping proportion is used to segment the raw data. Some of the samples contain null values which are removed from the data set. Five hundred samples are therefore present in each segment. Each windowed set of data is used to calculate feature vectors.

C. Automatic Feature Learning

The sample size for our studies is 500 × 6. We design a 6-head CNN to learn local features on each dimension of sensors. 6-head CNN is used to learn features from the x, y, and z-axis of the accelerometer and the x, y, and z-axis of the gyroscope, respectively. Each head in the proposed multihead CNN uses a different filter size to enable the network to learn discriminative features from the input time-series data and improve the performance. Hence, the input vector at time t is denoted by

$$S_t = \left[S_t^{\mathrm{acc}_x}, S_t^{\mathrm{acc}_y}, S_t^{\mathrm{acc}_z}, S_t^{\mathrm{gyro}_x}, S_t^{\mathrm{gyro}_y}, S_t^{\mathrm{gyro}_z}\right]. \tag{1}$$

In order to learn discriminant features, a 6-head CNN is designed to process the input vector as shown in Fig. 1. The convolutional layer utilizes a set of learnable kernels to perform the convolution operation. Using the activation function, the convolution operation produces the feature map for the next layer. The jth feature map at the lth layer of the hth head of the multihead CNN is denoted by $\alpha_{lj}^{i,h}$ which can be represented as

$$\alpha_{lj}^{i,h} = f\left(f_{\mathrm{conv2D}}^{h}\left(\alpha_{l-1}^{i+q}\right)\right), \text{ where h} = 1, 2, 3, \ldots, 6. \tag{2}$$

Here, $f(.)$ is the ReLU activation function used to reduce the vanishing gradient problem in the network. Negative input values become zero at the activation layer. i is the row number of the feature map matrix. The hth head convolution function of the proposed multihead CNN is denoted by $f_{\mathrm{conv2D}}^{h}(\alpha_{l-1}^{i+q})$ and can be represented as

$$f_{\mathrm{conv2D}}^{h}\left(\alpha_{l-1}^{i+q}\right) = b_{lj} + \sum_{p}\sum_{q=0}^{k_l^h-1} \omega_{ljp}^{q,h}\alpha_{(l-1)p}^{i+q,h} \tag{3}$$

where b_{lj} refer to the bias for this particular feature map, $\omega_{ljp}^{q,h}$ is the weight matrix at the position q of the convolutional kernel at layer l, k_l^h is the length of the kernel of the hth head at the lth layer, p is the index of the feature maps at the $(l-1)$th layer. multihead CNN derives local features with time dependency from sensory data and then fed to the next step for further processing.

D. Attention Mechanism

The attention mechanism was initially developed to focus on a particular area of an image rather than the entire picture to identify it. However, the emphasis changes over time. By modifying the weights of the data extracted by the multihead CNN, we adopt an attention network based on this concept to concentrate only on the critical features. Filtering out the significant representations for recognition is the task of the attention layer. The weights of representations are redistributed using an attention method. As opposed to single-head attention, multihead attention can better attend to more information while incurring a similar cost in computation, allowing the network to learn more quickly in parallel. In order to create a new intelligent activity recognition method, we think about combining this attention with CNN; more specific model parameters will be provided in the following section. The attention mechanism uses the "key-value" pair format to represent the input feature to map a query. The "key" is used to calculate the attention distribution and the "value" is used to generate the selected feature. Every key is attended to by the attention operator, which also calculates a similarity score that is then used to determine the weights for each value vector. Subsequently, we use the scaled dot-product to get the similarity score. Then we use softmax to get the attention weights. The values are then scaled according to their corresponding attention weight. The entire process is represented as

$$\mathit{Attention}(Q, K, V) = \mathit{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where $\mathit{softmax}(\alpha_{ij}) = [\exp(\alpha_{ij})/\sum_j \exp(X_j i)]$, Q is the query matrix, V is the value matrix and K is the key matrix, and d_k is the dimension of Q and k. For each query, the values associated with the keys with the highest similarity score are given a higher weight (i.e., higher importance). To put it another way, the weights are employed to give greater importance to values that are more relevant to the specific query. When the query, key, and value matrices together relate to elements in the same sequence, we refer to self-attention. Here, a multihead convolutional attention mechanism is used to exploit the features extracted using different convolution channels. Multihead attention mechanism executed parallel attention function c times. In our case, the value of c is 30. The multihead attention is represented as

$$\mathit{Multihead}(Q, K, V) = \mathit{Concat}(\mathit{head}_1, \mathit{head}_2, \ldots, \mathit{head}_c)W^O \tag{5}$$

where $\mathit{head}_i = \mathit{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W_i^Q, W_i^K, W_i^V, and W^O are the parameter matrices with dimension d_k/h, d_k/h, d_v/h, and d_O, respectively. In our experiment,

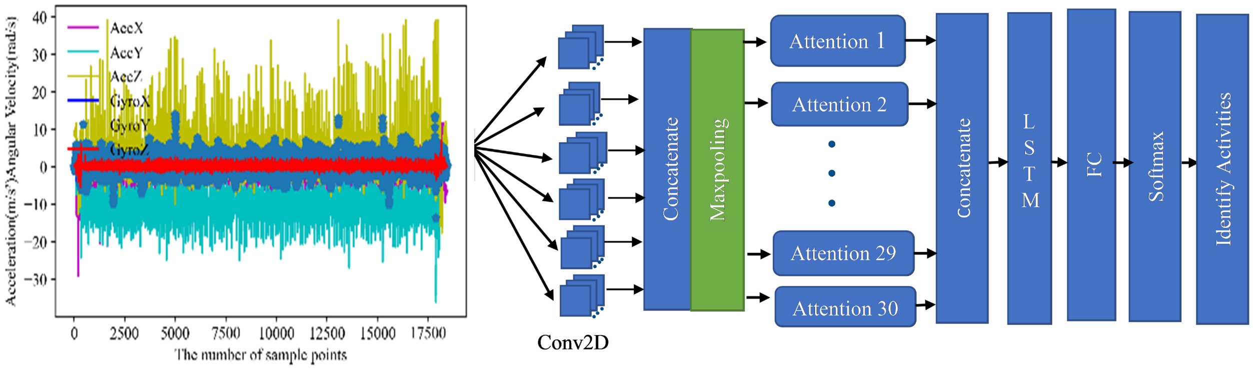

Fig. 1. Proposed attention-based deep learning architecture.

$d_k/h = d_v/h = 128$. The outputs of all the heads are then combined together.

E. Proposed Framework for HAR

Fig. 1 demonstrates the suggested HAR framework with an attention mechanism. To learn local features with time dependency, a sliding window of the 6-D sensory data is fed into multihead CNN. In order to keep the important details of the features from one sensor dimension and to decrease the parameters, the output of each of the six heads is concatenated and passes through a max-pooling layer. The attention mechanism then kicks in to modify the weights of the traits in order to keep the crucial features. These crucial elements are then sent into an LSTM network in order to develop latent feature representations. LSTM integrates the crucial features of each time segment into a unified sequential feature as follows:

$$h_t = f\big(w\big[h_{t-1}, x_t\big] + b\big) * \tanh(C_t) \tag{6}$$

where h_{t-1} is the hidden state of the previous time, x_t is the current time input, w is the weight of gate of LSTM cell, C_t is the current time memory unit, and h_t is the forward output.

To obtain more abstract characteristics, the output of the LSTM is processed through a fully connected layer (FCL). The final step is to categorize physical activities using a softmax layer. In conclusion, we develop an attention-based multihead CNN followed by LSTM that can automatically learn important features from sensory data gathered at high-sampling rates instead of depending on feature engineering.

IV. Performance Evaluation

In this section, we first describe the specifics of the data set used in the studies. The basic descriptions of the experimental setup employed in this work are then given. Finally, we display the outcomes of our suggested method's experimental testing.

A. Data Set

We conduct our experiments using the UCI public standard data set, as described in Section III-A. This data collection contains 2738449 samples of a total of six different activities, including Bike (A1), Walk (A2), Stand (A3), Sit (A4), Stairsup (A5), and Stairsdown (A6). Table II is a list of the specific details for this data set.

TABLE II
Instances of Each Activity in UCI Data Set

Activities	Instances	Percentage
Bike (A1)	542729	19.8%
Walk (A2)	510014	18.6%
Stand (A3)	442724	16.2%
Sit (A4)	419883	15.3%
Stairsup(A5)	415320	15.2%
Stairsdown(A6)	407779	14.9%
Total	2,738,449	100%

B. Experimental Setup

To validate the proposed method, we compare it with several complex learning algorithms, including the "random forest" (RF), "support vector machine" (SVM), "extreme learning machine" (ELM), "artificial neural network" (ANN), and "multilayer perceptron" (MLP), as well as the DL algorithms, 1-D-CNN, 2-D-CNN, LSTM, "recurrent convolutional neural network" (RCNN), and "gate recurrent unit" (GRU). The approaches mentioned above are either used for HAR or combined CNN modules with recurrent modules. The hyperparameters involved in the approaches mentioned above and the proposed solution are fine-tuned using the 10-fold cross-validation (CV) on the training set. A grid search is applied to calculate the hidden nodes in ELM, ANN, and MLP with the validation set. Grid search is also applied to optimize the hyperparameters of SVM. The best hyperparameter values are C = 1000, γ = 0.001, and kernel = "rbf." Five hundred decision trees are used in RF for ensemble learning.

The data set is separated into two groups to carry out the experiment: 30% of the volunteers are chosen to test the suggested HAR solution, while 70% of the volunteers are chosen for training. Therefore, neither the training nor the testing use data from the same subjects. Since CV is less computationally intensive, we employ it in our experiment to create several splits of the training and validation sets from the training set. For a more thorough analysis, a leave-one-subject-out (LOSO) CV is also carried out. In this, data from one subject are used for testing, and data from the other subjects are used for training. The fact that the test data is concealed from the models makes this cross-subject test more rigorous. It provides a more realistic environment for assessing the generalization

TABLE III
RECOGNITION PERFORMANCE OF ALL THE ACTIVITIES

Activities	Precision	Recall	F1-score
A1	96.10%	96.34%	96.22%
A2	97.53%	98.10%	97.81%
A3	95.13%	96.12%	95.62%
A4	97.12%	97.39%	97.27%
A5	96.08%	96.70%	96.39%
A6	95.90%	96.98%	96.44%

capabilities of the models. 2-D convolution is performed in the input layer of the multihead CNN using a smartphone-based tri-axial accelerometer and gyroscope data. The 3-axial accelerometer and 3-axial gyroscope take six different heads in the multihead CNN. In this experiment, the convolution layers with kernel size 3 and stride 2 are activated using the "ReLU" activation function. In the max-pooling layer, the stride and pooling sizes are both of size 2. The rate of learning is set at 0.001. Three different filter sizes are used in the different heads of the multihead CNN. For head-1 and head-4, the filter size is 3×3, for head-2 and head-5, the filter size is 5×3, and for head-3 and head-6, the filter size is 7×3. The deep LSTM comprises one LSTM layer with a size of 64, an FCL layer with a size of 128, and a softmax layer for classification. The suggested fusion structure uses one LSTM layer with a size of 64. In order to approach or achieve the optimal value, the Adam optimizer calculates and adjusts the network parameters that have an impact on the model's training and output, decreasing the loss function.

C. Experimental Results for Offline Recognition

The purpose of the offline recognition is to implement and test the proposed framework in a PC for activity recognition with the previously collected data and then the offline implemented model can be deployed on a smart device to test the model for real-time, online data. This enables a user-independent, device-independent HAMR system for activity monitoring and recognition. We present the experimental findings of the suggested technique in this part in terms of accuracy, precision, recall, and ROC score. The findings of our suggested method for identifying each activity are examined next, and they have condensed in Fig. 2 and Table III. Fig. 2 shows the confusion matrix on test data using our proposed model. The misclassification rate of Sit(A3), and Stand(A4) activities are higher in comparison with other activities. It may be possible due to the static nature of these activities. Table III shows the average precision, recall and F1-score for the test data using 10-fold CV. Fig. 3 demonstrates the accuracy and loss with respect to the number of epochs with our proposed model. The loss curve starts to converge just after 40 epochs as shown in Fig. 3.

1) Comparison With Traditional Approaches: The aforementioned traditional ML and DL-based approaches are tested using the same smartwatch-based UCI data set with 10-fold CV and the experimental results (accuracy, F1-score, and ROC-AUC score) are tabulated in Table IV. Our proposed approach achieves higher performance with the accuracy and F1-score of 97.07% and 96.30%, respectively. Hence, our proposed method in which Multihead CNN with attention mechanism followed by LSTM learns the features more efficiently to achieve higher performance. Table V demonstrates the performance of different methods using LOSO CV. This cross-subject test is more challenging since the models are not aware of the test data, which creates a more realistic environment for testing the generalization skills of the models [4]. According to the findings shown in Tables IV and V, our suggested strategy performs better than the aforementioned methods when employing both the 10-fold and LOSO CV.

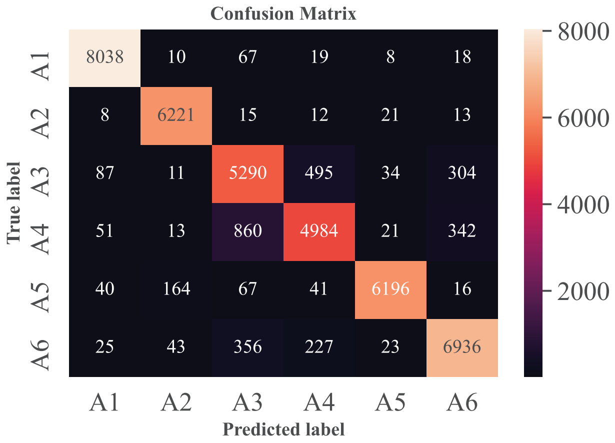

Fig. 2. Confusion matrix on test data of the proposed method.

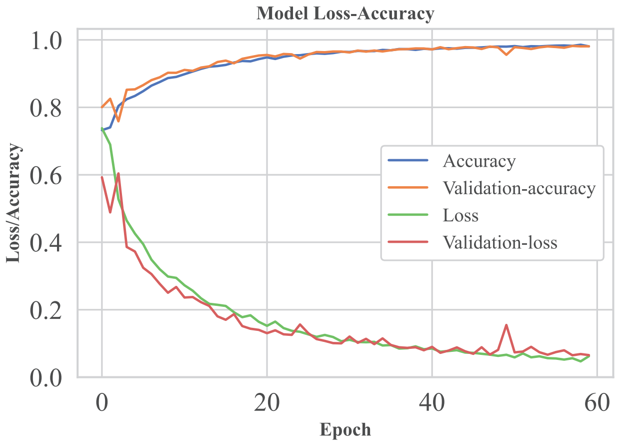

Fig. 3. Accuracy and loss w.r.t epochs of the proposed method on test data.

TABLE IV
EXPERIMENTAL RESULTS OF DIFFERENT METHODS USING 10-FOLD CV

Methods	Accuracy	F1-Score	ROC-AUC Score
ANN	78.12%	0.77	0.85
SVM	82.67%	0.82	0.91
ELM	81.45%	0.80	0.87
MLP	76.43%	0.76	0.85
RF	91.17%	0.92	0.98
1D-CNN	89.61%	0.90	0.97
2D-CNN	91.86%	0.93	0.98
LSTM	88.26%	0.90	0.96
RCNN	92.43%	0.93	0.99
GRU	93.45%	0.92	0.99
Proposed Method	97.07%	0.96	1.00

2) Ablation Experiment: Ablation is a long-used technique in neuroscience that involves inflicting controlled damages

TABLE V
EXPERIMENTAL RESULTS OF DIFFERENT METHODS USING LOSO CV

Methods	95% Confidence Interval	True Accuracy
ANN	(0.7491, 0.7516)	75.09%
SVM	(0.7912, 0.7989)	79.45%
ELM	(0.7723, 0.7806)	77.87%
MLP	(0.7319, 0.7378)	73.43%
RF	(0.8911, 0.9005)	89.32%
1D-CNN	(0.8622, 0.8686)	86.39%
2D-CNN	(0.8913, 0.8977)	89.47%
LSTM	(0.8523, 0.8609)	85.89%
RCNN	(0.9038, 0.9096)	90.72%
GRU	(0.9018, 0.9104)	90.79%
Proposed Method	(0.9598, 0.9659)	**96.03%** & **95.77%**(F1-Score)

TABLE VI
EXPERIMENTAL RESULTS OF ABLATION STUDY

Methods	Attention	Accuracy	F1-score	ROC
CNN	-	88.56%	0.871	0.91
CNN	Yes	90.12%	0.887	0.92
LSTM	-	89.23%	0.882	0.92
LSTM	Yes	91.07%	0.910	0.93
CNN-LSTM	No	90.14%	0.901	0.94
CNN-LSTM	Yes	93.17%	0.920	0.96
Multihead CNN	No	93.62%	0.934	0.96
Multihead CNN	Yes	94.06%	0.938	0.97
Multihead CNN-LSTM	No	95.24%	0.937	0.96
Proposed Method	Yes	97.07%	0.966	1.00

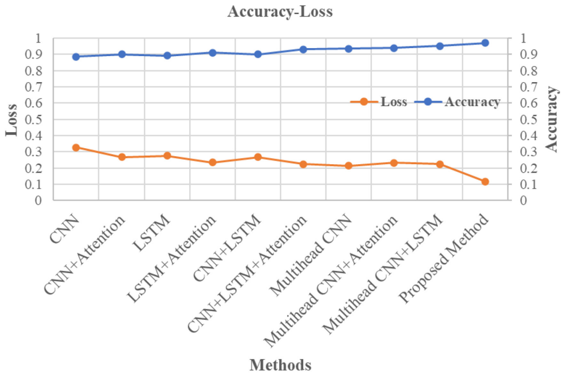

Fig. 4. Accuracy and loss for different variations of the proposed method.

on neural tissue to see how they affect the brain's ability to accomplish given tasks. This method yields an in-depth understanding and explanations of each aspect of the activity's pattern and function in response to external stimuli [30]. Ablation is used to better comprehend DL-based approaches as a natural extension. An ablation study examines the performance of a system by deleting specific components to determine the component's contribution to the system. An ablation study as shown in Table VI is performed to reveal the importance of each component such as multihead, CNN, attention, and LSTM included in the proposed method. First, we experiment only using the CNN model without an attention network. The designed CNN learns the sequential local features on the data streams and achieves an accuracy of 88.56% to detect activities. As LSTM has strong sequential modeling capabilities, we use only the LSTM model without an attention network and achieve an accuracy of 89.23%, which is greater than the CNN model. After that, we combine both the CNN and LSTM models to recognize activities. The integrated CNN-LSTM, according to Table VI, increases model performance. The integrated CNN-LSTM with the attention layer also enhances the performance of the model. To see the effectiveness of six different heads with different filter sizes, we use multihead CNN with and without an attention layer. Multihead CNN with attention layer gives higher accuracy in comparison with the previous approaches, as shown in Table VI. Then, we combine LSTM with CNN to encode the temporal dependencies of the learned local features and, subsequently, high-level representation. After combining CNN-LSTM, we also adopted an attention network to dynamically adjust the importance of the features extracted using multihead CNN, the results further improved. In Fig. 4, the results on test data are listed. The proposed model still has the best performance thinking about accuracy and loss at the same time. Its accuracy is 97.07%, which is a little higher than different variations of the models, but the loss is far better than these. Thus, Multihead CNN, attention, and LSTM are used and added to our proposed model.

3) Comparison of the Proposed Method With SOTA Approaches: In the HAR literature, several attention-based deep learning methods are proposed, which are presented in Section II. To compare our proposed method's effectiveness with other current methodologies described in the literature, mentioned in Table VII, we employ accuracy and F1-score as the performance metric. The mentioned benchmark schemes used either a smartwatch or smartphone-based accelerometer and gyroscope data to demonstrate the performance of their proposed model. Out of the five benchmark schemes, the attention mechanism was used in [7], [8], and [23]. The achieved accuracy and F1-score in [7] and [23], are lower than our proposed method. In [8], the achieved accuracy is higher than our proposed method using the WISDM [22] data set but lesser using the UCI [20] data set. Khan and Ahmad [8] set the batch size to 130 and the number of training epochs to 260, which is very high compared to our proposed method. Moreover, the WISDM data set is based on a tri-axial accelerometer only. Further, in [8] the experimental results were not evaluated using either 10-fold CV or LOSO. Mekruksavanich and Jitpattanakul [31] used WISDM smartwatch data to show the performance of their proposed model. Ni et al. [21] proposed a multihead CNN-LSTM without attention and got an accuracy and F1-score of 98%, which is more than our proposed model just because they have used the smartphone-based UCI HAR data set. This data set is well-processed and consists of 561 statistical features. Moreover, in [21], bidirectional LSTM has been used, which is slow to train. Hence, not suitable for real-time applications. In our proposed method, we use the raw smartwatch-based sensory signal collected using two different models of smartwatches with different sampling frequencies to train the model. Without the attention layer using our proposed Multihead CNN-LST architecture, we get 95.24% accuracy and 93.72% F1-score, respectively. The proposed attention mechanism dynamically adjusts the importance of features from the two modalities. Thus, we can say that our proposed model outperforms the existing benchmark classifiers yielding an increased F1-score rate of 96.3% with the attention layer.

TABLE VII
COMPARISON OF THE PROPOSED METHOD WITH SOTA METHODS

Reference	Method	Dataset	Accuracy	F1-score
[7]	Multihead Convolutional Attention	WISDM [22]	96.4%	95.4%
[8]	Attention induced Multi-head CNN	WISDM[22]	98.18%	97.20%
		UCI[20]	95.38%	95.37%
[23]	TRASEND	UCI-WISDM [24]	-	84.8%
[31]	Hybrid LSTM	Smartwatch [32]	96.2%	96.3%
[21]	Multi-head CNN-LSTM	UCI[20]	98%	98%
[25]	CNN-based Bi-LSTM parallel model with attention	UCI[20]	96.7%	-
		WISDM[22]	95.86%	-
[28]	HAP-DNN	DaLiAc[29]	94.55%	-
	Multi-head CNN-LSTM	UCI-WISDM [24]	95.24%	93.72%
	Proposed Method	UCI-WISDM [24]	97.07%	96.30%

TABLE VIII
PERFORMANCE OF THE PROPOSED MODEL FOR THE REAL-TIME, ONLINE EXPERIMENT ON SMARTPHONE

Device	Activity Subject	Sit	Stand	Jog	Walk	Upstairs	Downstairs	Bike	Average
Smartwatch	Subject1	97.2%	96.1%	-	96.6%	95.7%	96.1%	94.7%	96%
	Subject5	98.3%	96.9%	-	96.7%	95.2%	95.9%	96.8%	96.6%
	Subject7	98.8%	97.2%	-	97.3%	96.9%	96.6%	97.9%	97.5%
Smartphone	Subject1	95.7%	93.1%	92.9%	94.6%	93.7%	92.9%	-	93.8%
	Subject2	96.4%	94.2%	93.8%	94.7%	93.9%	94.3%	-	94.6%
	Subject3	96.8%	95.1%	94.5%	95.6%	93.8%	94.7%	-	95.1%

D. *Real-Time, Online Experiment on Smartdevice*

HAR requires efforts to build a generalized model using the training data sets with the hope to achieve good performance in test data sets. However, in real applications, the training and testing data sets may have different distributions due to various reasons, such as different body shapes, acting styles, and habits, damaging the model's generalization performance. Hence, to validate the generalization of the proposed model, an unseen, independent smartphone-based data set is collected to test the performance of the same with one of the different activities, i.e., jogging. The proposed activity recognition framework is also implemented for Android mobile using an Android programming language. Previously, for offline activity recognition, 70% of the 2738449 instances of the collected data, presented in Section III-A, are used for training. Whereas, due to the limited computation power of smartphones, ten instances of each activity for six subjects (subjects 2, 3, 4, 6, 8, and 9) are chosen from the previously selected training data to train the on-device model. The test data is taken from the rest of the subjects (subjects 1, 5, and 7) whose activity data is not involved in training data. The purpose of this online model is to monitor the real-time activity of the user. In this experiment, data preprocessing (segmentation) and classification are performed on the Android phone itself. To prove the generalization of the proposed model, it is also tested on the online data stream, which was collected using a smartphone instead of a smartwatch. The smartphone-based collected data is new, unseen, and independent of the trained model. Three different subjects (subjects 1, 2, and 3) data, while keeping the mobile in the front pant pocket or hand, are tested on the smartphone. The selected activities are "sitting," "standing," "walking," "jogging," "walking upstairs," "walking downstairs," and "biking." Among all these activities, biking data is collected using a smartwatch, and jogging data are collected using a smartphone only. In both cases, the average classification accuracy is higher. This signifies that the proposed framework is device independent as well as user independent.

There is no significant difference in the average accuracy for all the activities performed by different users with various device-based unseen, independent data with different positions, as shown in Table VIII. This makes the proposed model more efficient and intelligent as it overcomes user-dependent, device-dependent, and position-dependent issues. This application is also tested on a Samsung M32 mobile with MediaTek Helio G80 processor to check the CPU utilization. It uses under 8% CPU utilization. Hence, this application can also be run with other applications simultaneously.

V. CONCLUSION AND FUTURE DIRECTIONS

This article proposes a novel attention-based deep learning approach to HAR based on a tri-axial accelerometer and gyroscope signal in the context of IoHT. In this work, we address the following major challenges: 1) using training data sets from two different smartwatches with two different tri-axial sensors; 2) the use of the proposed method in a smartphone i.e., a memory and energy constraints device; and 3) validate the generalizability of the proposed architecture using an unseen, independent, real-time data set. The proposed method combines multihead CNN, an attention mechanism, and an LSTM network. The proposed framework does not require specific data preprocessing and feature engineering methods which are mandatory in traditional HAR approaches. The proposed framework automatically learns important discriminant features using an attention mechanism to efficiently recognize human activities. The attention mechanism works as a mitigation technique to overcome data heterogeneity impairments. The results of the experiments show that the suggested method may get a very high-identification rate of 97.07% using 10-fold CV and 96.03% using LOSO CV when testing accuracy is measured and F1-score of 96.30% and 95.77% using 10-fold and LOSO CV, respectively. Moreover, the proposed method is tested using unseen and independent data from both the smartphone and smartwatch are 95.5% and 96.7%, respectively, which validates the generalizability of the proposed model.

There are several future directions and research challenges concerning next-generation computing using DL. In line with the recent study of [33], future work will consider edge computing with a higher number of activities and even under more complex situations as new challenges. Moreover, the developed algorithm will be validated with several wearable devices to improve its identification ability for tiny devices with limited memory, computational power, and energy [34], [35]. We also expect to embed intelligence into the proposed model for real-time decision making.

References

[1] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, M. Ryan, M. Elhoseny, and H. Song, "ST-DeepHAR: Deep learning model for human activity recognition in IoHT applications," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4969–4979, Mar. 2021.

[2] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep spatial–temporal model based cross-scene action recognition using commodity WiFi," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3592–3601, Apr. 2020.

[3] D. Thakur, A. Guzzo, and G. Fortino, "T-SNE and PCA in ensemble learning based human activity recognition with smartwatch," in *Proc. IEEE 2nd Int. Conf. Human-Mach. Syst. (ICHMS)*, 2021, pp. 1–6.

[4] D. Thakur, S. Biswas, E. S. L. Ho, and S. Chattopadhyay, "ConvAE-LSTM: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition," *IEEE Access*, vol. 10, pp. 4137–4156, 2022.

[5] Z. Chen, M. Wu, W. Cui, C. Liu, and X. Li, "An attention based CNN-LSTM approach for sleep-wake detection with heterogeneous sensors," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3270–3277, Sep. 2021.

[6] D. Thakur and S. Biswas, "Attention-based deep learning framework for hemiplegic gait prediction with smartphone sensors," *IEEE Sensors J.*, vol. 22, no. 12, pp. 11979–11988, Jun. 2022.

[7] H. Zhang, Z. Xiao, J. Wang, F. Li, and E. Szczerbicki, "A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1072–1080, Feb. 2020.

[8] 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.

[9] T. Zebin, M. Sperrin, N. Peek, and A. J. Casson, "Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 1–4.

[10] S. Yu and L. Qin, "Human activity recognition with smartphone inertial sensors using Bidir-LSTM networks," in *Proc. 3rd Int. Conf. Mech., Control Comput. Eng.*, 2018, pp. 219–224.

[11] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual Bidir-LSTM for human activity recognition using wearable sensors," *Math. Problems Eng.*, vol. 2018, Dec. 2018, Art. no. 7316954.

[12] W. Ye, J. Cheng, F. Yang, and Y. Xu, "Two-stream convolutional network for improving activity recognition using convolutional long short-term memory networks," *IEEE Access*, vol. 7, pp. 67772–67780, 2019.

[13] H. Wang et al., "Wearable sensor-based human activity recognition using hybrid deep learning techniques," *Security Commun. Netw.*, vol. 2020, p. 12, Jul. 2020.

[14] R. Mutegeki and D. S. Han, "A CNN-LSTM approach to human activity recognition," in *Proc. Int. Conf. Artif. Intell. Inf. Commun.*, 2020, pp. 362–366.

[15] G. Ercolano and S. Rossi, "Combining CNN and LSTM for activity of daily living recognition with a 3D matrix skeleton representation," *Intell. Service Robot.*, vol. 14, no. 2, pp. 175–185, Apr. 2021.

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

[17] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.

[18] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Netw.*, vol. 116, pp. 237–245, Aug. 2019.

[19] W. Ahmad, B. M. Kazmi, and H. Ali, "Human activity recognition using multi-head CNN followed by LSTM," in *Proc. 15th Int. Conf. Emerg. Technol.*, 2019, pp. 1–6.

[20] 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. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2013, pp. 1–6.

[21] C. Ni, S. Guan, and Y. Li, "Human activity recognition using a improved model based on multi-head CNN-LSTM," in *Proc. 7th Int. Conf. Inf. Sci. Control Eng.*, 2020, pp. 688–693.

[22] 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.

[23] 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.

[24] A. Stisen et al., "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *Proc. 13th ACM Conf. Embedded Netw. Sens. Syst.*, 2015, pp. 127–140.

[25] X. Yin, Z. Liu, D. Liu, and X. Ren, "A novel CNN-based bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data," *Sci. Rep.*, vol. 12, no. 1, p. 7878, May 2022.

[26] I. D. Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable sensor-based human activity recognition with transformer model," *Sensors*, vol. 22, no. 5, p. 1911, 2022.

[27] 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.

[28] Y. Zhou, Z. Yang, X. Zhang, and Y. Wang, "A hybrid attention-based deep neural network for simultaneous multi-sensor pruning and human activity recognition," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25363–25372, Dec. 2022.

[29] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, multi-sensor based classification of daily life activities: Comparison with state-of-the-art algorithms using a benchmark dataset," *PLoS One*, vol. 8, no. 10, 2013, Art. no. e75196.

[30] P. H. Schiller, "The effect of superior colliculus ablation on saccades elicited by cortical stimulation," *Brain Res.*, vol. 122, no. 1, pp. 154–156, 1977.

[31] S. Mekruksavanich and A. Jitpattanakul, "Smartwatch-based human activity recognition using hybrid LSTM network," in *Proc. IEEE SENSORS*, 2020, pp. 1–4.

[32] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[33] S. S. Gill et al., "AI for next generation computing: Emerging trends and future directions," *Internet Things*, vol. 19, Aug. 2022, Art. no. 100514.

[34] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 1, pp. 115–133, Jan. 2013.

[35] S. Qiu et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Inf. Fusion*, vol. 80, pp. 241–265, Apr. 2022.

Open Access funding provided by 'Università della Calabria' within the CRUI CARE Agreement