Information Fusion 94 (2023) 17–31

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things

Md. Milon Islam [a,*], Sheikh Nooruddin [a], Fakhri Karray [a,b], Ghulam Muhammad [c]

[a] *Centre for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, N2L 3G1, Ontario, Canada*
[b] *Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates*
[c] *Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia*

ARTICLE INFO

Keywords:
Human activity recognition
Multi-head convolutional neural network
Convolutional block attention module
Convolutional long short term memory
Internet of things

ABSTRACT

Human Activity Recognition (HAR) has become a crucial element for smart healthcare applications due to the fast adoption of wearable sensors and mobile technologies. Most of the existing human activity recognition frameworks deal with a single modality of data that degrades the reliability and recognition accuracy of the system for heterogeneous data sources. In this article, we propose a multi-level feature fusion technique for multimodal human activity recognition using multi-head Convolutional Neural Network (CNN) with Convolution Block Attention Module (CBAM) to process the visual data and Convolutional Long Short Term Memory (ConvLSTM) for dealing with the time-sensitive multi-source sensor information. The architecture is developed to be able to analyze and retrieve channel and spatial dimension features through the use of three branches of CNN along with CBAM for visual information. The ConvLSTM network is designed to capture temporal features from the multiple sensors' time-series data for efficient activity recognition. An open-access multimodal HAR dataset named UP-Fall detection dataset is utilized in experiments and evaluations to measure the performance of the developed fusion architecture. Finally, we deployed an Internet of Things (IoT) system to test the proposed fusion network in real-world smart healthcare application scenarios. The findings from the experimental results reveal that the developed multimodal HAR framework surpasses the existing state-of-the-art methods in terms of multiple performance metrics.

1. Introduction

Human Activity Recognition (HAR) refers to the recognition of movement behavioral patterns by analyzing human activity signals, which has a significant economic value and scientific research importance in the area of smart healthcare [1]. Real-time monitoring of the activities of the elderly is a key task in the development of smart health monitoring that employs signals from wearable and mobile devices to enhance medical decision support systems [2,3]. Hence, HAR has evolved into a dynamic and peremptory research area for the effective detection of human activities and interactions in ambient assisted living environments, which is being widely investigated for the aptly called Internet of Healthcare Things (IoHT) [4,5].

Nowadays, there are numerous wearable sensors integrated with intelligent devices including smartphones and smartwatches, as well as numerous wearable components that can be attached to multiple on-body positions to collect and transfer real-time activity data through wireless sensor networks [6,7]. These sensing paradigms offer effective data fusion for efficient HAR that supports the development of services and applications for the elderly in smart healthcare frameworks based on the generation of real-time fused information [8,9]. Over the years, several human activity recognition frameworks have been introduced based on various approaches including sensors [10], cameras [11], Wi-Fi signals [12], and others. The sensors are more useful for HAR because they are more reliable in testing situations, do not need proper lighting, and preserve human privacy. However, the sensor-based HAR systems are expensive and burdensome for their users. The major advantages of vision-based HAR systems are that they spare regular users from having to wear a number of cumbersome devices on various parts of their bodies. Nevertheless, these systems have privacy concerns, and their efficiency is influenced by the level of ambient lighting. On the contrary, Wi-Fi based HAR techniques use signal descriptors to identify the human activities that are reflected in the mobility of Wi-Fi signals. However, these techniques have a limited range, require significant investment, and entail unified protocols as well as benchmarks.

[*] Corresponding author.
E-mail addresses: milonislam@uwaterloo.ca (M.M. Islam), sheikh.nooruddin@uwaterloo.ca (S. Nooruddin), karray@uwaterloo.ca (F. Karray), ghulam@ksu.edu.sa (G. Muhammad).

https://doi.org/10.1016/j.inffus.2023.01.015
Received 12 August 2022; Received in revised form 10 January 2023; Accepted 13 January 2023
Available online 20 January 2023

Conventional Machine Learning (ML) techniques [13] are widely used for HAR in smart healthcare applications. The major task of ML approaches includes the extraction of the most significant features from the raw data. In general, time-domain and frequency-domain attributes are exploited for ML-based HAR. The derivative features are fed into ML algorithms for human activity recognition purposes. Several machine learning-based human activity recognition frameworks [14,15] have been developed to address the challenges associated with real-time HAR in IoHT. However, the hand-crafted features that need subject-matter expertise and human experience present some serious challenges for ML algorithms. Deep Learning (DL) algorithms [16] have recently been introduced for HAR purposes as they have demonstrated their exceptional efficiency in automatic feature extraction as well as recognition. Deep learning technique [17] extracts the necessary features from the multimodal raw data that are quite effective to enhance the performance of HAR compared with conventional techniques. In literature, various human activity recognition systems have been proposed based on deep learning approaches. The most popular and extensively exploited approaches for HAR are Convolutional Neural Networks (CNNs) [18], which use several convolutional kernels to extract hidden patterns from raw visual data. To handle sequential data, Recurrent Neural Networks (RNNs) [19,20] have shown promising performance for long-term dependency on multi-sensor information. Besides, the time-series data collected from wearable sensors are encoded to image representation including recurrence plot, Gramian angular field, and Markov transition field in recent studies [21]. Moreover, the current researches are highly focused on attention mechanism with deep learning architectures for HAR applications in IoHT environment. For many time-series and visual data classification applications [22], the attention mechanism has become essential and performs better than traditional methods. Attention mechanisms focus on the most significant extracted attributes and remove the unnecessary noisy attributes during the recognition — resulting in overall performance improvement.

As multiple wearable devices are utilized to develop human activity recognition frameworks, it is required to synchronize and combine the various sensor information using a uniform data fusion approach to record more complicated activities from and multimodal and multipositional perspective. Keeping this in mind, the researchers are now introducing various fusion strategies to develop multimodal HAR. In most of the works, single nature data (time-series) collected from accelerometer, and gyroscope are utilized for developing HAR based on fusion techniques [23]. The research gaps exist in developing fusion approaches for multi-nature data such as fusing the information of multi-sensory (time-series) and camera (visual) for multimodal HAR.

This paper proposes a multi-level feature fusion approach for multimodal HAR that efficiently processes and recognizes human activities recorded by multiple sensors and stationary cameras embedded in a variety of Internet of Things (IoT) devices. Considering the multiple sensor information along with visual data recorded from various wearable sensors, a fusion approach is developed to enhance the performance of HAR in smart healthcare applications. The proposed system consists of several blocks for multimodal human activity recognition. The multi-head CNN with Convolution Block Attention Module (CBAM) extracts the channel information and spatial information from visual data. A Channel Attention Module (CAM) and Spatial Attention Module (SAM) are added to each head of CNN to extract features in the channel and spatial dimensions. The extracted features from each head are fused to get all the attributes of the visual data. To handle the sequential data in the form of time-series collected by multiple sensors, Convolutional Long Short Term Memory (ConvLSTM) architecture is utilized that extracts the temporal attributes from raw time-series data. Lastly, the output of these two blocks is fused and sent to a fully connected layer through the Global Average Pooling (GAP) layer for recognition purposes. The proposed multi-level feature fusion approach is evaluated with an open-access large multimodal dataset called UP-Fall detection dataset and it was compared with two baseline models along with existing HAR frameworks to verify its performance. Besides, the proposed framework is deployed in IoHT environment to facilitate real-time human activity recognition. To the best of our knowledge, this is the first article that applies multi-head CNNs integrated with Convolution Block Attention Module to concurrently perform channel and spatial feature fusion of visual information based on the existing state-of-the-art literatures in this area. Additionally, we first used all the data including time-series and images of this dataset for our experiments.

The key contributions of this article are given below.

(1) A fusion technique is introduced for multimodal human activity recognition that can fuse the information of cameras and multiple sensors simultaneously through the fusion of the output of multi-head CNNs with ConvLSTM in IoHT environment.
(2) An efficient feature fusion approach is developed to retrieve the channel and spatial dimension attributes using CBAM and fusing the output of each head of CNN architecture.
(3) A ConvLSTM-based architecture is developed with a multi-sensor based data fusion method to handle sequential data and retrieve the high-level temporal features from the raw signals.
(4) We compare the experimental findings of the proposed fusion architecture with two baseline models along with the state-of-the-art systems on a large multimodal dataset called UP-Fall detection dataset to verify the performance of our developed architecture.
(5) We compared the model performances on two different IoT deployment scenarios.

The remaining parts of the article are structured as follows. Section 2 presents the relevant literature of the most recent HAR systems in the view of single-modal and multimodal data using deep learning techniques for smart healthcare applications. Section 3 demonstrates the proposed fusion architecture incorporating the multi-head CNN with CBAM, ConvLSTM, IoT deployment, and evaluation measures. Section 4 demonstrates the experimental specifications, the details of the dataset, the experimental findings, the comparison with the state-of-the-art, and ablation studies. Lastly, the conclusion, limitations, and potential future works are discussed in Section 5.

2. Related works

In recent years, wearable technologies are in ongoing development stage incorporating complicated and sensitive embedded sensors. In general, DL and ML algorithms are utilized to improve the efficiency of HAR using data retrieved from these sensors. This section describes the recent developments of human activity recognition platforms in the area of DL, ML, and IoT. Here, we discussed a few most recent HAR approaches based on the single-modal and multimodal data.

2.1. Single-modal based HAR

Single-modal based human activity recognition systems use only single source data for recognition purposes. Single-modal based HAR utilizes any sensor data such as accelerometer, gyroscope, or RGB camera [24]. The development of the single-modal based HAR platform is quite cost-effective but it lacks reliability and robustness.

Al-qaness et al. [25] proposed a multilevel residual network with attention called Multi-ResAtt for HAR from wearable sensors data. The system consists of a variety of parallel-oriented residual modules and initial blocks. A Bidirectional Gated Recurrent Unit (BiGRU) with attention blocks are preceded by two dense layers for HAR in this proposed system. The developed scheme is tested by three publicly available datasets and the best accuracy of 90.08% is found from the PAMAP2 dataset. Lu et al. [26] developed a wearable prototype to capture human activities from the environment through the use of one tri-axial accelerometer for HAR. The system extracted global and local features from the data to find out the effect of different perspectives of human activities. The extracted features are fed into several ML techniques to develop classification models and an average accuracy of 96% is obtained from Random Forest leveraging both global and local features. Zhou et al. [27] introduced a semi-supervised learning architecture exploiting the concept of Long Short Term Memory (LSTM) with deep Q-network to enrich the performance of human activity recognition on weakly labeled sensor information. Deep Q-network automatically classifies the data with a reward based on distance to address the issue of lack of labeled data. The LSTM network recognized the fine-grained features contextually retrieved from sequential motion information. The experimental findings reveal that the presented scheme achieved a Receiver Operating Characteristic (ROC) curve up to 0.95.

In another research, Abdel-Basset et al. [28] presented a lightweight DL architecture for HAR utilizing the data from the wearable heterogeneous sensors. The collected sensor data is encoded into RGB images and the features are extracted using a hierarchical multi-scale extraction module. The evaluations are performed with two open-access datasets and the accuracy of 98% and 99% are obtained from University of California Irvine (UCI) HAR and Mobile HEALTH (MHEALTH) datasets respectively. However, the proposed framework did not mention about power consumption and inference latency in IoT environment. Further, Abdel-Basset et al. [29] used Convolutional Neural Network, LSTM, and attention techniques to develop a dual-channel network called ST-DeepHAR for human activity recognition. In this system, an adaptive channel-squeezing technique is introduced to fine-tune the feature extraction capacity of CNN by using multichannel dependency. The proposed framework is evaluated with two publicly available datasets and obtained the best accuracy of 98.9% from the Wireless Sensor Data Mining (WISDM) dataset. Zhang et al. [30] developed a human activity recognition platform based on multi-head techniques along with attention mechanisms. The high-level features are retrieved from the raw data exploiting multi-head CNN and fused to generate a single feature vector. Thirty parallel attention heads are exploited to select the most significant attributes for HAR. Besides, the system appraised an F-measure of 0.954 using the WISDM dataset with the parameters of 2.77 million. However, the complexity of the proposed network is relatively high.

2.2. Multimodal based HAR

Multimodal based HAR frameworks use multiple sensors to record real-world environmental data. Multimodal based HAR systems are merely deployed to classify Activities of Daily Living (ADLs) from a variety of data sources including sensors, and visuals [31]. The multimodal based frameworks depend on a fusion of sensors including accelerometers, gyroscopes, or depth cameras for data collection purposes.

Yadav et al. [32] proposed an activity recognition framework called ARFDNet utilizing pose estimation based classification architecture. The skeleton features of the individuals are extracted from the RGB videos using the pose estimation model. The skeleton features are preprocessed and fed into CNNs preceded by Gated Recurrent Units (GRUs) to learn the spatiotemporal dynamics. Finally, the classification is done in a fully connected layer using the outcome of the GRUs. The experimental findings revealed that the developed network obtained the best accuracy of 96.7% from the UP-Fall detection dataset. However, the system used only data from camera 2 (frontal view). Ramirez et al. [33] demonstrated a HAR system that used human skeleton features from raw images collected from a standard video camera. The scheme is able to detect the human falls as well as different types of activities for several individuals in the same scene through the use of human skeleton features. The system applied four different machine learning algorithms and the average accuracy of 98.59% is found from Camera 1 (lateral view) of the UP-Fall detection dataset. Inturi et al. [34] introduced a deep learning architecture that is able to detect different kinds of human activities and falls. The system used human joint points that are captured by applying the AlphaPose pre-trained architecture. The retrieved key points are processed using CNNs to analyze the spatial correlation of the key points. The LSTM network preserved the long-term dependencies in this system. UP-Fall detection dataset is utilized to prove the efficiency of the developed architecture. An accuracy of 98.59% is achieved from the experimental results with the camera data only.

In another study, Lin et al. [35] proposed an adaptive multimodal fusion architecture for HAR using videos and inertial data. The skeleton sequence patterns are retrieved through the proposed Spatio-temporal graph CNN with adaptive loss function and the inertial data features are extracted using LSTM with a fully convolutional network. The contribution of two modalities at the decision level are learnt using the proposed adaptive learning technique. The architecture is tested with the data called H-MHAD collected from a laboratory environment along with two open-access datasets UTD Multimodal Human Action Dataset (UTD-MHAD) and Continuous Multimodal Human Action Dataset (C-MHAD). Moreover, the best accuracy of 91.18% is found from the C-MHAD dataset. Ranieri et al. [36] used a two-stream ConvNet for HAR improved with LSTM and a temporal CNN to analyze the temporal data on videos and inertial sensing devices. Here, the authors deployed two fusion strategies such as feature-level and late fusion to enrich the recognition performances. The experiments have been carried out with two publicly available popular datasets including egocentric multimodal and UTD-MHAD. The proposed network obtained the best accuracy of 85.47% from the UTD-MHAD dataset. Further, Ranieri et al. [37] presented a HAR platform utilizing different types of data such as videos, ambient sensors, and inertial units in an ambient assisted living environment. The main focus of this research is to generate a multimodal dataset called Heriot-Watt University/University of Sao Paulo (HWU-USP) activities dataset in a small kitchen. As a DL framework, the system used CNN and LSTM for recognition purposes and achieved an accuracy of 98.61% from the experiments. The framework is also evaluated with the UTD-MHAD dataset and found an accuracy of 92.33%. Furthermore, Gao et al. [38] developed a framework for recognizing human action that incorporated latent information from various viewpoints. The system proposed a category-level dictionary learning architecture based on adaptive fusion. The design of query sets and the development of the regularization scheme for the assignment of the adaptive weight were performed in order to incorporate dictionary learning. This method found excellent results, with an increase in accuracy of between 3% to 10%.

3. Proposed fusion approach for multimodal HAR

In this article, we propose a multi-level feature fusion architecture for multimodal human recognition in smart healthcare applications. The proposed system consists of several components such as the end IoT devices, deep learning-based fusion architecture, cloud server, Fifth-Generation (5G) communication, and stakeholders. Fig. 1 shows the overall system architecture for multimodal HAR in IoHT. In our system, the end IoT devices include accelerometer, gyroscope, camera, EEG headset, infrared sensor, and ambient light sensor to collect the different ADLs in visual and time-series modalities. The fusion architecture is developed to be able to handle multimodal signals (visual and time-series) using multi-head CNN along with CBAM as well as ConvLSTM. The proposed DL model is deployed in the cloud server using large training data samples. The parameters of the fusion network are stored in the cloud to facilitate smart healthcare applications. All the data transfers from end IoT devices to the cloud server as well as the cloud server to stakeholders are done through Wi-Fi signals. The decision of the proposed fusion architecture is conveyed to the stakeholders through 5G communication so that they can take further actions.

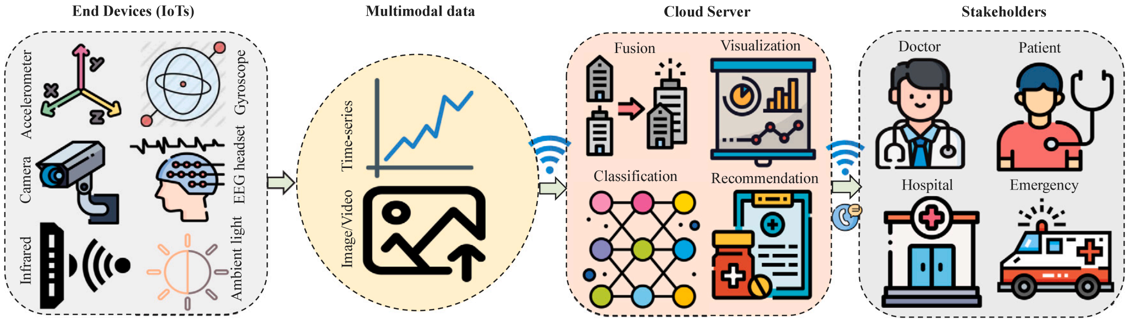


Fig. 1. System architecture of the developed multimodal HAR system in Internet of Healthcare Things. Here, the cameras and multiple sensors collect the human daily activities in the shape of visual and time-series. The deep learning-based fusion network is developed and deployed in the cloud server. The notification is sent to the stakeholders through 5G communication.

The proposed multi-level feature fusion architecture contains several blocks to recognize human actives from the multimodal data. Fig. 2 illustrates the several blocks of the proposed architecture for multi-modal human activity recognition for smart healthcare applications. The data collected from multiple wearable sensors in the form of time-series are passed to ConvLSTM architecture to obtain the high-level temporal features. The visual information from cameras are sent to three-head CNN networks where each contains the CBAM network to extract the patterns from the channel and spatial dimensions. In this level, the outputs (homogeneous features) of each head are fused to get the high-level features from the raw images. In the next level, the extracted features from visual data from multi-head network and multi-sensor information from ConvLSTM architecture are fused to get the final features and are sent to fully connected layer through the GAP layer. The GAP layer generates one feature map for each class of human activities. Finally, the classification module containing fully connected layer with Softmax function recognizes the human activities. The following sections demonstrate the proposed fusion architecture with IoT deployment and its blocks in detail.

3.1. Multi-head Convolutional Neural Network

CNN [39] has become one of the most representative neural networks in the area of DL for image recognition. In this paper, a three-head Convolutional Neural Network (as illustrated in Fig. 2) is designed to retrieve the relevant patterns from the input images. The most pivotal component in CNN is the convolutional layer [40] that contains several convolutional filters which produce the output feature map from the input images through convolution operation. In the convolution layers, the output feature maps from the previous layer are convolved using a number of convolutional kernels. Moreover, a bias is used to enhance the output of the convolution operation that is passed through the activation function to generate the feature map for the following layer. Mathematically, the qth feature map at the pth layer of hth head of multi-head CNN is a matrix, and the value at the xth row is represented as $z_{pq}^{x,h}$. The value is calculated as in (1).

$$z_{pq}^{x,h} = f_{ReLU}\left(f_{conv2d}^{h}\left(z_{p-1}^{x+l}\right)\right) \qquad \forall_h = 1,2,3 \tag{1}$$

Here, f_{ReLU} is the activation function that substitutes all negative values with zero in the feature map, and f_{conv2d}^{h} represents convolution function of the hth head in our multi-head CNN, as expressed in (2).

$$f_{conv2d}^{h}\left(z_{p-1}^{x+l}\right) = b_{pq} + \sum_{r} \sum_{l=0}^{n_p^h-1} w_{pqr}^{l,h} z_{(p-1)m}^{x+l,h} \tag{2}$$

Here, the bias for a specific feature map is b_{pq}, r is the index of the feature maps at the $(p-1)$th layer, $w_{pqr}^{l,h}$ is the weight matrix at the position l of the convolution kernels, and n_p^h is the length of the kernel of the hth head in our multi-head CNN.

Another important component in our proposed multi-head CNN architecture is the pooling layer [41] that minimizes the number of parameters and calculations by scaling down the spatial size of the feature description. The most popular and widely used pooling technique is called max pooling, which selects the largest component within each receptive area as shown in (3).

$$\mathcal{Y}_{kpq} = max_{(a,b)\in\mathbb{R}_{p,q}} \mathcal{V}_{kab} \tag{3}$$

Here, $\mathcal{Y}_{kpq}$ denotes the pooling operation of kth feature maps, and $\mathcal{V}_{kab}$ stands for the component at position (a,b) enclosed by pooling region $\mathbb{R}_{p,q}$ that represents a receptive field around the location (p,q).

In the proposed multi-level feature fusion architecture, two convolution operations with filter sizes of 32, and 64 are utilized to retrieve the high-level patterns from the input images. The size of the input data is 64 × 64 in this network. We consider the kernel size of 3 × 3, 5 × 5, and 7 × 7 in head1, head2, and head3, respectively. Three max-pooling operations with pool size 2 × 2 are exploited to minimize the number of parameters and calculations. A dropout rate of 0.3, 0.3, and 0.4 are put after each max-pooling operation to avoid over-fitting during the training process.

3.2. Convolutional Block Attention Module

In the proposed architecture, each head contains two CBAMs (as shown in Fig. 2) to enhance the training performance by highlighting the channel and spatial features of the activity images. The CAM network enables the proposed architecture to focus on important channel features and ignores the other features. To measure the significance of each channel, various weight information is applied to diverse feature dimensions and feature channels of the visual data. The SAM network allows the proposed architecture to give more attention to the spatial dimension information on the feature map. For feature extraction, CBAM [42] sequentially extracts a 1D channel attention map $X_c \in \mathbb{R}^{C\times1\times1}$ and a 2D spatial attention map $X_s \in \mathbb{R}^{1\times H\times W}$ from the given intermediate feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$ of the activity visual data. The overall attention mechanism is given in (4).

$$\begin{aligned} \mathcal{F}' &= X_c(\mathcal{F}) \otimes \mathcal{F} \\ \mathcal{F}'' &= X_s(\mathcal{F}') \otimes \mathcal{F}' \end{aligned} \tag{4}$$

Here, ⊗ presents the element-wise multiplication, and $\mathcal{F}''$ is the final refined feature. During the multiplication, the channel attention features are compressed along the spatial dimension, and vice-versa.

The CAM network (as shown in Fig. 2) enhances the weight of the relevant information and suppresses the weight of the unnecessary information in the feature channel. Thus, the proposed architecture highlights more on the discriminative channels in the activity images. In the CAM network, the average-pooled patterns, and the max-pooled attributes are extracted from the aggregated feature map of spatial information by exploiting both average-pooling and max-pooling operations. The extracted high-level patterns are sent to a shared Multi-Layer

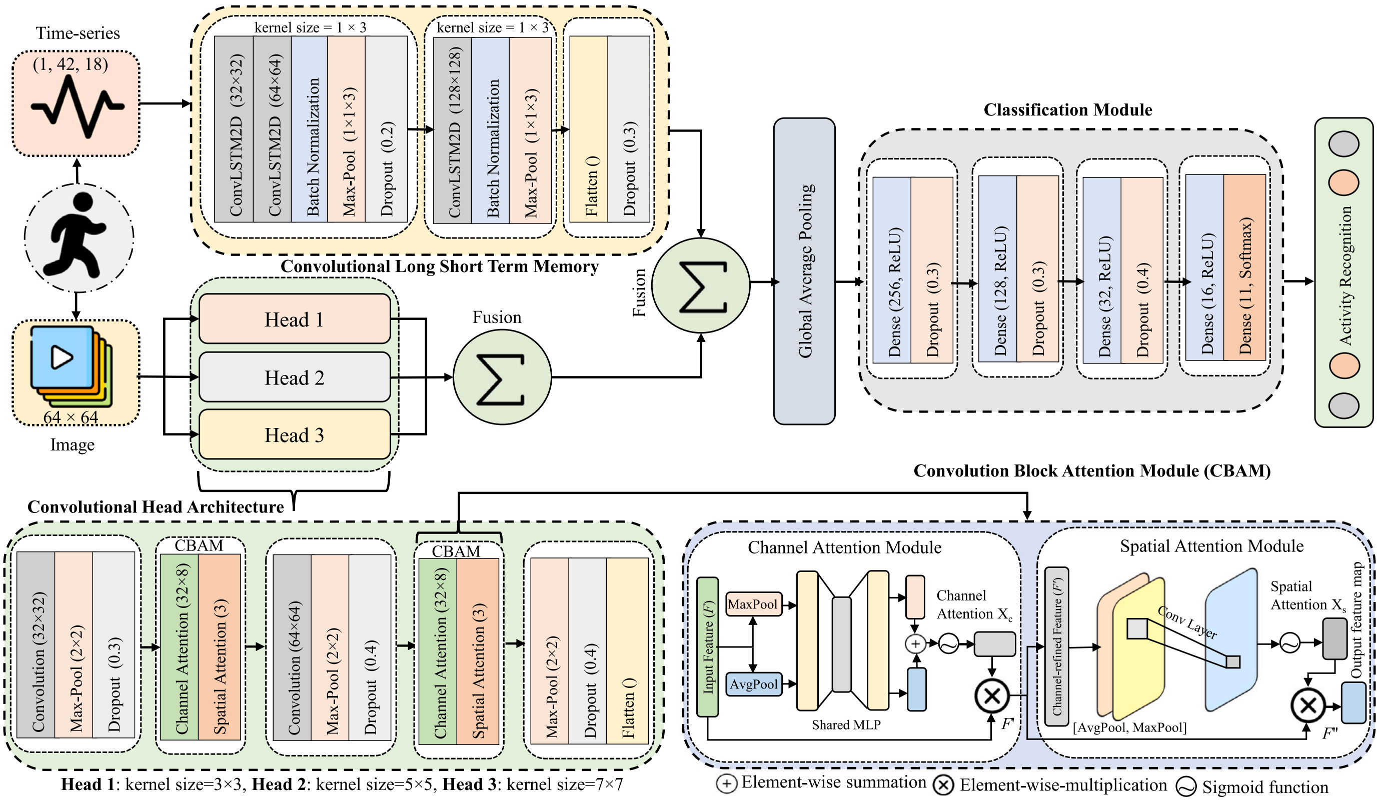


Fig. 2. Schematic diagram of the developed fusion architecture for multimodal human activity recognition in a smart healthcare framework. The time-series data recorded from multi-sensors are fed into ConvLSTM architecture to get the temporal features (see top part). A three-head CNN architecture with CBAM extracts the high-level channel and spatial dimension features from the visual data and the outputs of each head are get fused (see lower part). Additionally, the output features of two-branch networks are fused and passed to the classification module through the GAP layer. Finally, the Softmax function is utilized to provide output for HAR.

Perceptron (MLP) model that contains one hidden layer. Moreover, the output of the shared network is forwarded through a pipeline containing further max-pooling and average-pooling operations, and a non-linear activation function ReLU to generate channel attention map $X_c \in \mathbb{R}^{C\times 1\times 1}$. The use of two pooling operations facilitates to extract high-level features efficiently. The channel attention is mathematically calculated as in (5).

$$X_c(\mathcal{F}) = \sigma(MLP(AvgPool(\mathcal{F}))) + (MLP(MaxPool(\mathcal{F}))) \tag{5}$$

Here, σ represents the sigmoid function.

The SAM network (as depicted in Fig. 2) improves the spatial dimension features in the feature map by performing feature filtering on the pixels at various positions in the same spatial dimension and weighting the significant features. The SAM performs the max-pooling and average-pooling operations on the given feature map F' along the channel dimension to get two feature maps. The retrieved features are fused and convolved by a convolution layer with a kernel size of 7×7 to generate the final spatial attention map $X_s \in \mathbb{R}^{1\times H\times W}$. The mathematical formulation of the SAM is given in (6).

$$X_s(\mathcal{F}') = \sigma(f^{7\times 7}([AvgPool(\mathcal{F}'); MaxPool(\mathcal{F}')])) \tag{6}$$

Here, $f^{7\times 7}$ is the convolution operation with a filter size of 7×7.

3.3. Convolutional Long Short Term Memory

Convolutional Long Short Term Memory architectures [43] are an improvement of the CNN, RNN, and LSTM networks. RNN networks transmit historical information through chain network structures. However, as the distances between the chain network structures become larger, the long-term information becomes harder for the RNN networks to learn. LSTM networks improve this condition of RNN via adding three different gate functions that control memory information. However, LSTM's use fully connected networks which leads to information redundancy. LSTM networks only keep temporal dependencies but do not take spatial dependencies into account.

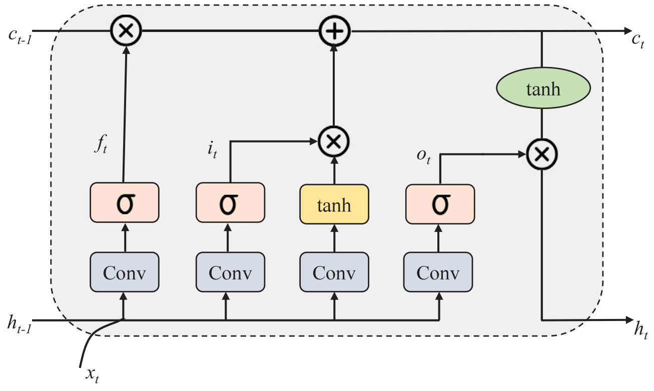


Fig. 3. The architecture of ConvLSTM. The new memory c_t and output h_t are formed by updating the internal memory c_{t-1} based on the current input x_t and the prior output h_{t-1}.

CNNs are generally a combination of convolutional layers and pooling layers that can extract spatial information from input data modalities but cannot extract temporal dependencies. ConvLSTM networks embed convolution operations inside LSTM cells, thus alleviating the redundancy of fully connected layers and resulting in fewer parameters compared to LSTM networks due to weight sharing features of convolutional kernels. ConvLSTM networks can extract both the spatial and temporal dependencies from input data modalities as they combine both CNN and LSTM's. ConvLSTM networks have better spatiotemporal comprehension ability compared to fully connected LSTM networks. The architecture of ConvLSTM is shown in Fig. 3. ConvLSTM networks can be expressed using the following major operations.

ConvLSTM networks have three kinds of gates: input, forget, and output. The input gates can be formulated using (7).

$$i_t = \sigma(W_{xi} * x_{(i)} + W_{hi} * h_{(t-1)} + W_{ci} \odot c_{(t-1)} + b_i) \tag{7}$$

Here, the input features are denoted by $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. The cell outputs are represented as $c_{(1)}, c_{(2)}, \dots, c_{(n)}$. The hidden cells are denoted as $h_{(1)}, h_{(2)}, \dots, h_{(n)}$. The input, forget, and output gates are expressed by $i_{(t)}, f_{(t)}$, and $o_{(t)}$. $W_{x\sim}$ and $W_{h\sim}$ represents the 2D convolutional kernels. The convolution operation in all subsequent representations is expressed by $*$. The Hamdard product is expressed by $\odot$.

Eq. (8) represents the formulation of the forget gates.

$$f_{(t)} = \sigma(W_{xf} * x_{(t)} + W_{hf} * h_{(t-1)} + W_{cf} \odot c_{(t-1)} + b_f) \tag{8}$$

The cell states are calculated using the input and forget gates. Eq. (9) presents the formulation for calculating the cell states.

$$c_{(t)} = f_{(t)} \odot c_{(t-1)} + i_{(t)} \odot \tanh(W_{xc} * x_{(t)} + W_{hc} * h_{(t-1)} + b_c) \tag{9}$$

The mathematical expression of the output gates require the cell states. The formulation for calculating the output gates is shown in (10)

$$o_{(t)} = \sigma(W_{xo} * x_{(t)} + W_{ho} * h_{(t-1)} + W_{co} \odot c_{(t)} + b_o) \tag{10}$$

Eq. (11) formulates the calculation of the hidden states. The cells, hidden states, and three gates are considered as 3D tensors.

$$h_{(t)} = o_{(t)} \odot \tanh(c_{(t)}) \tag{11}$$

In our proposed fusion architecture, the input shape for the ConvLSTM stream is (1,42,18). The input contains data collected from multiple types of sensors in the shape of time-series. Three layers of ConvLSTM were used inside the stream with filter sizes of 32, 64, and 128. In all the ConvLSTM layers, the kernel size was 1 × 3 and Rectified Linear Unit (ReLU) was used as the activation function. We used Batch normalization, max-pooling layers, and dropout layers to ensure the model does not overfit or underfit the input data. The max-pooling layer has a kernel size of 1 × 1 × 3. We used 20% and 30% dropout in the dropout layers.

3.4. Classification module

The output vectors from ConvLSTM network and multi-head CNN architecture are fused and sent into the classification module to predict the target classes through the global average pooling layer. The GAP layer creates one feature map of each class of human activities and minimizes the number of trainable parameters. The fully connected layers contain the neurons of 256, 128, 32, and 16 with ReLU activation function, and the fully connected layers are followed by dropout layers to prevent over-fitting. Finally, the Softmax activation function calculates the class score of each activity and outputs the correct activity class that obtained the higher probability score. The prediction representation of the Softmax function is formulated as in (12) and (13), where, $\mathcal{F}$ presents the output features from the previous fully connected layer and a is the number of activities in the data samples (N).

$$\mathcal{P} = SoftMax(\mathcal{F}) = \frac{exp(\mathcal{F})}{\int_1^a exp(\mathcal{F})} \tag{12}$$

$$\hat{y} = argmax(\mathcal{P}) \tag{13}$$

Moreover, the cross-entropy loss was used to reduce the loss value during training. The loss function ($\mathcal{L}$) is demonstrated as in (14), where, y_i is the actual activities, and $\hat{y}_i$ is the predicted activities obtained by the proposed architecture.

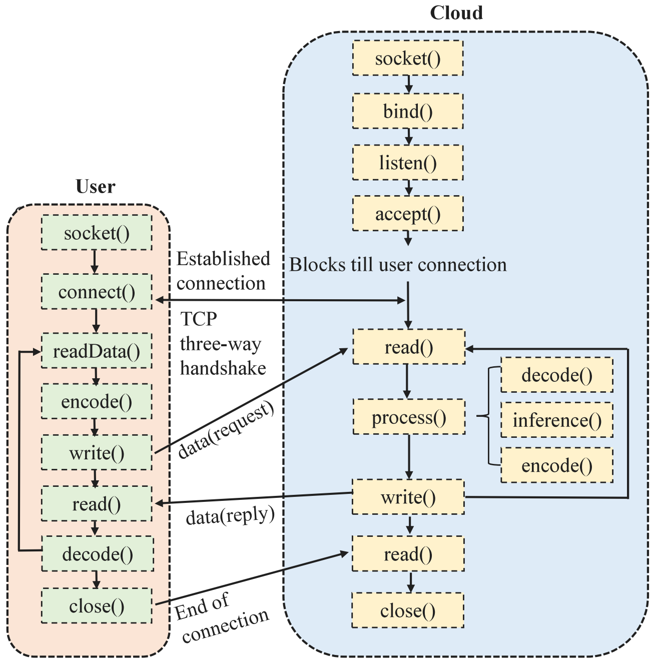


Fig. 4. Block diagram of user application and cloud server application for interaction for multimodal human activity recognition in IoHT.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{14}$$

3.5. IoT deployment

Once the fusion architecture is trained and evaluated, it is ready to be deployed as an IoT application. In this scenario, the proposed system would be deployed in the cloud. Individual user applications can then connect to the cloud server that is serving the model, send data to the server, and receive inference results. The first possible deployment network scenario is where the cloud server and users are in the same network. In this case, the latency between the users and server would be minimal. This scenario is feasible for large nursing homes, hospitals, and care facilities that want to provide localized monitoring services to their customers. The second possible deployment network scenario is where the cloud server and users are in different networks. This scenario is more feasible for corporations that want to provide Deep Learning as a Service (DLaaS). Although the latency between the cloud server and user application would be much larger than in the same network scenario, the different network configuration is more scalable. The cloud server would be more sophisticated, have specialized system architecture, have load balancing, and redundancy to ensure minimal downtime. Fig. 4 provides a block diagram of the processes involved in the interaction between the user applications and cloud service. The connection and data transfer requires sockets. Generally, the cloud service has a fixed Internet Protocol (IP) and port number. The user applications must know the fixed IP address and port number to connect to the cloud service. The internet protocol used in the network connection and data transfer is Transmission Control Protocol (TCP). TCP provides three-way handshake, retransmission in case of transmission failure, error detection, network congestion avoidance, and improved security over other internet protocols such as User Datagram Protocol (UDP). TCP protocol has slightly higher latency than UDP. However, TCP provides much more security-oriented features than UDP. After a connection is established, the client reads data continuously from the sensors, encodes them, and sends them via the network to the cloud server. The server receives the data, decodes it, and performs inference. The server then encodes the inference results, sends it to client, and waits for further data. The client decodes the inference results, takes action based on the predicted activity, and starts sending the next batch of data. The network connection can be closed at any time from the client or server. The diagram shown in Fig. 4 depicts a single-thread network. When the number of clients and servers will be large, multithreading and threadpool executor blocks would be required to efficiently complete the tasks.

3.6. Performance evaluation

In this article, we measure the efficiency of the developed fusion architecture in terms of four widely used statistical metrics: accuracy, precision, recall, and F1-Score. Additionally, to evaluate the performance of the fusion architecture for multimodal human activity recognition in IoHT environment, we measure two metrics including latency of the deployed IoT network and response time of the proposed fusion network. The lower value of these parameters represents the better performance of the system.

- **Latency**: Latency refers to the time taken for a packet or data message to reach the destination from its point of origin. The low latency network ensures real-time applications with minimal delay times.
- **Response time**: Response time includes latency as well as the time taken for the server to perform inference on the received data. It is mathematically demonstrated as follows.

$$Response\ time = 2 \times latency + processing\ time \tag{15}$$

4. Experiments results and discussions

In this section, we described the experimental findings of our developed fusion technique for multimodal HAR to show the efficacy of the proposed system. This section includes a brief description of the experimental setup, the multimodal dataset for the experiments, the quantitative results analysis, and the ablation study. Moreover, a comparative study is conducted to compare our findings with two baseline models as well as recent state-of-the-art methods.

4.1. Experimental setup

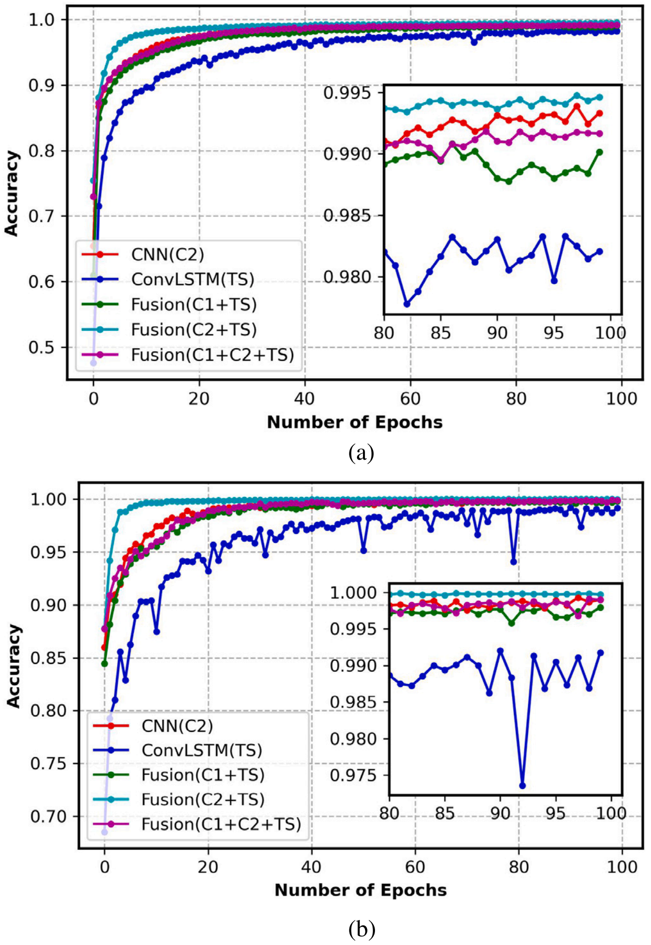
The developed multi-level feature fusion architecture for multimodal HAR was written in Python (v2.10) and implemented with the Keras API on top of the TensorFlow (v2.8.0) backend. The training was conducted in a desktop computer equipped with GPU: GTX 1070 with 8 GB GDDR5 memory, CPU: Intel® Core™ i7-6700K CPU @ 4.00 GHz × 8, and RAM: 32 GB DDR4. The computer was running in an Ubuntu operating system with a version of 18.04 LTS.

4.2. Dataset description

We used a large and publicly available multimodal dataset called UP-Fall detection dataset [44] for our experimentation. To collect the ADLs data, 17 healthy participants (9 males and 8 females) were considered between 18 to 24 years, with an average height and weight of 1.66 m and 66.8 kg respectively. The dataset comprises 11 human activities including 6 basic ADLs and 5 different kinds of fall activities. The human activities and falls along with the duration of each activity stored in this dataset are given in Table 1. The duration of all fall events is the same (60 s) and the duration of daily activities is different. Each activity in the dataset was recorded 3 trials by each healthy participant. The data collected from the wearable sensors and vision devices are in the shape of time-series and visual respectively.

Table 1
Description of the activities performed by human participants in UP-Fall detection dataset.

Activity number	Description	Duration (s)
$A1$	Falling forward using hands	10
$A2$	Falling forward using knees	10
$A3$	Falling backwards	10
$A4$	Falling sideward	10
$A5$	Falling sitting in empty chair	10
$A6$	Walking	60
$A7$	Standing	60
$A8$	Sitting	60
$A9$	Picking up an object	10
$A10$	Jumping	30
$A11$	Laying	60

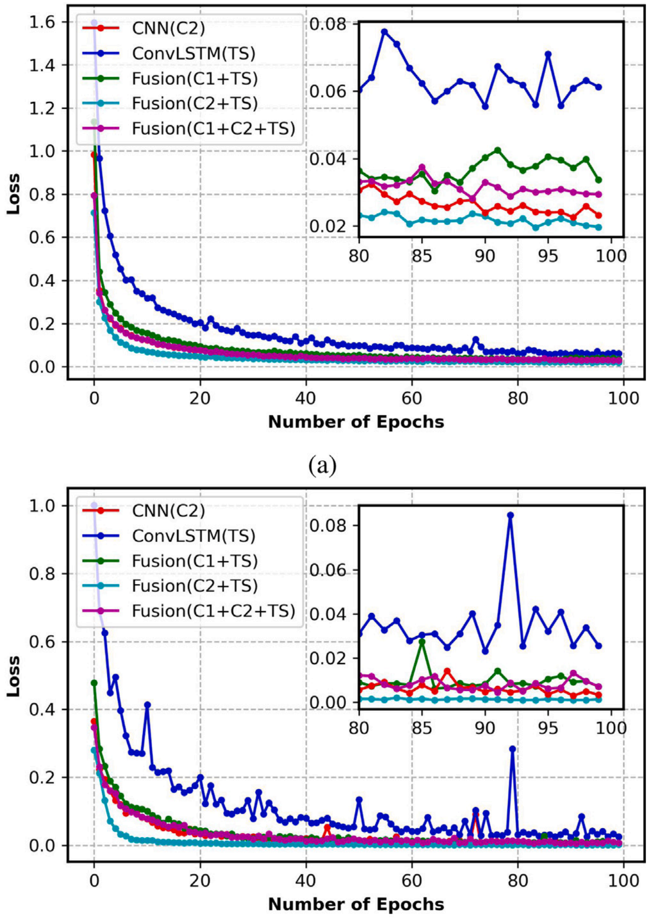


(a)

(b)

Fig. 5. Performance measures (accuracy curves) of the developed fusion approach along with two baseline architectures. (a) Training (b) Validation.

In our research, we used all modalities of data to conduct the experiments. We considered trial 1 as well as trial 2 of each activity as the training set and trial 3 of each activity as the testing set. The image data is resized to 64 × 64 from 640 × 480 to feed into the network. The data collected from wearable sensors are in time-series format and the total number of attributes is 42. We generated the sequences from the raw time-series data to process them using ConvLSTM. In the experiment, the sequence size is 18 depending on the smallest number of data points.



(a)

(b)

Fig. 6. Performance measures (loss curves) of the developed fusion approach along with two baseline architectures. (a) Training (b) Validation.

4.3. Quantitative results analysis

We conducted the experiments for 100 epochs to train and evaluate the proposed multi-level feature fusion architecture. We performed Grid Search analysis to select the best hyperparameters for our architecture. Based on Grid Search results, the learning rate is set to 0.001, the optimizer is Adam, the batch size is 128, and the loss function is sparse categorical cross-entropy. In the experiments, two baseline architectures such as CNN and ConvLSTM are considered to get the results from the visual and time-series data individually. Moreover, we run three variants of our proposed fusion architecture based on the nature of the dataset. The CNN ($C2$) is run only with the visual data from camera 2 as the previous studies [32] depicted that the existing architectures showed robust performance for the frontal view of the camera data. The time-series data is fed into ConvLSTM (TS) to achieve the relevant attributes from the raw multi-sensor information as well as the performance of this architecture for time-series data. Three variants of fusion architecture are Fusion ($C1 + TS$), Fusion ($C2 + TS$), and Fusion ($C1$+$C2$+TS) where the $C1$, $C2$, and TS represent the data from camera1 (lateral view), camera 2 (frontal view), and time-series data from multi-sensor, respectively. Besides, we utilized 20% of the training dataset as a validation set to validate the proposed fusion architecture. In the experimental studies, we presented the average performance of ten runs while the weights of the architectures are randomly initialized during each run. We also depicted the best performance from the output of the run that scores the highest performance.

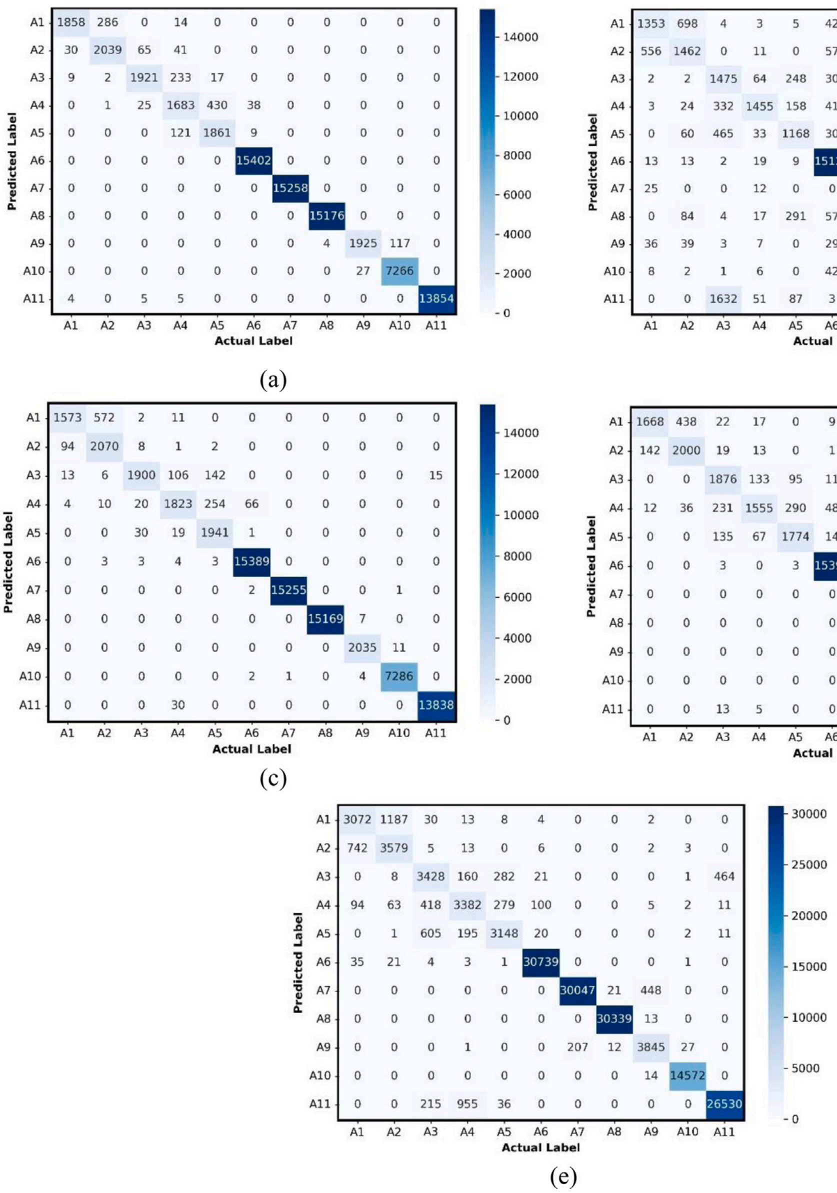
Fig. 5 depicts the relationship between the number of epochs and training as well as validation accuracies of our proposed multimodal human activity recognition architecture. From Fig. 5, it is evident that the training processes stop at nearly 100 epochs. It is observed from the experimental findings that the baseline model ConvLSTM takes more time to generalize as it contains the time-series data only. Another baseline model CNN achieves comparatively better performance compared to ConvLSTM as it has experimented with only visual data. Our fusion architecture containing the visual data from camera 2 (frontal view) along with all time-series data (42 attributes) is generalized well in both training and validation phases compared to two other use cases of our fusion architectures such as fusion for camera 1 data with time-series information and fusion for all data. More time is required for fusing all data as it comprises a relatively large amount of training data samples.

Besides, the loss curves are illustrated in Fig. 6 to show the relationship between epochs and losses (both training and validation) for five use cases. Similar to Fig. 5, the fusion approach containing data from camera 2 and time-series achieves fast convergence with a smaller number of epochs. The loss values of the other two use cases of fusion architecture such as Fusion ($C1 + TS$) and Fusion ($C1 + C2 + TS$) is relatively low but it is slightly higher than the Fusion ($C2 + TS$) architecture. The baseline model CNN obtains relatively low loss values compared to another baseline model ConvLSTM both in the training and validation set.

Fig. 7 represents the confusion matrices obtained by our proposed fusion architecture for multimodal human activity recognition. It is found from Fig. 7 that the fusion architecture with camera 2 data and time-series information recognizes the activities more efficiently than the other two fusion architecture variants. The baseline models misclassified a large number of activities compared to fusion architectures. It is observed from the fusion architecture, the activities $A1$ to $A5$ are achieving comparatively less scores than other activities. The highest recognition rate is obtained for the activity standing. The falling activities ($A1$ to $A5$) and the activity picking up an object ($A9$) which is just like a fall activity receive a relatively low score for activity recognition. In general, falling events consist of transitioning from standing to laying activities. In our dataset, a few of the first frames depict standing while a few of the final frames show laying. Hence, the proposed fusion architecture shows poor generalization for recognizing the fall activities. Moreover, more than 99% (0.99) of walking, standing, sitting, jumping, and laying are recognized while the lowest recognition rate of 73% (0.73) is obtained from the falling activity: falling forward using hands. In this architecture, falling forward using hands is misidentified as falling forward using knees as these behaviors are intimately associated and intertwined. Our fusion network learned that both activities are almost similar which demonstrates the higher error on the activity falling forward using hands.

Table 2 demonstrates the results of the developed fusion architecture along with baseline networks for UP-Fall detection dataset for each activity with respect to precision. Notably, the developed architecture discriminates the activities with more than 98% of precision for fusing the camera 2 data and time-series information. The lowest precision rate of 80.79% is obtained for activity $A2$ from CNN ($C2$), 48.20% for the activity $A3$ from ConvLSTM (TS), 74.70% for the activity $A3$ from Fusion ($C1 + TS$), 80.89% for the activity $A2$ from Fusion ($C2 + TS$), and 60.40% is achieved for the activity $A3$ from Fusion ($C1$+$C2$+TS). The lowest precision rate contains between activities $A2$ to $A3$. In some cases, the highest precision of 100% is found from the experiments for the activities $A6$–$A8$, and $A10$–$A11$ considering the baseline models as well as fusion architectures.

Similar to Table 2, the class-specific recall of the proposed fusion architecture along with baseline models are illustrated in Table 3. Here, the average recall rate of more than 98% is found from the experiments for Fusion ($C2 + TS$). The highest recall rate of 100% is achieved from the activities $A6$–$A8$, and $A11$ using CNN ($C2$), ConvLSTM (TS) architecture receives the best recall rate of 99.20%, Fusion ($C1 + TS$),



(a)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
A1	1858	286	0	14	0	0	0	0	0	0	0
A2	30	2039	65	41	0	0	0	0	0	0	0
A3	9	2	1921	233	17	0	0	0	0	0	0
A4	0	1	25	1683	430	38	0	0	0	0	0
A5	0	0	0	121	1861	9	0	0	0	0	0
A6	0	0	0	0	0	15402	0	0	0	0	0
A7	0	0	0	0	0	0	15258	0	0	0	0
A8	0	0	0	0	0	0	0	15176	0	0	0
A9	0	0	0	0	0	0	0	4	1925	117	0
A10	0	0	0	0	0	0	0	0	27	7266	0
A11	4	0	5	5	0	0	0	0	0	0	13854

(b)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
A1	1353	698	4	3	5	42	25	8	1	14	5
A2	556	1462	0	11	0	57	12	0	51	21	5
A3	2	2	1475	64	248	30	0	1	3	30	327
A4	3	24	332	1455	158	41	20	44	62	16	22
A5	0	60	465	33	1168	30	4	120	6	7	98
A6	13	13	2	19	9	15123	68	14	97	33	11
A7	25	0	0	12	0	0	14531	173	449	4	64
A8	0	84	4	17	291	57	212	14291	61	0	159
A9	36	39	3	7	0	29	346	61	1505	3	17
A10	8	2	1	6	0	42	0	0	1	7233	0
A11	0	0	1632	51	87	3	10	14	0	0	12071

(c)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
A1	1573	572	2	11	0	0	0	0	0	0	0
A2	94	2070	8	1	2	0	0	0	0	0	0
A3	13	6	1900	106	142	0	0	0	0	0	15
A4	4	10	20	1823	254	66	0	0	0	0	0
A5	0	0	30	19	1941	1	0	0	0	0	0
A6	0	3	3	4	3	15389	0	0	0	0	0
A7	0	0	0	0	0	2	15255	0	0	1	0
A8	0	0	0	0	0	0	0	15169	7	0	0
A9	0	0	0	0	0	0	0	0	2035	11	0
A10	0	0	0	0	0	2	1	0	4	7286	0
A11	0	0	0	30	0	0	0	0	0	0	13838

(d)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
A1	1668	438	22	17	0	9	0	0	4	0	0
A2	142	2000	19	13	0	1	0	0	0	0	0
A3	0	0	1876	133	95	11	0	0	0	0	67
A4	12	36	231	1555	290	48	0	0	5	0	0
A5	0	0	135	67	1774	14	0	0	0	0	1
A6	0	0	3	0	3	15396	0	0	0	0	0
A7	0	0	0	0	0	0	15134	40	83	1	0
A8	0	0	0	0	0	0	0	15175	1	0	0
A9	0	0	0	0	0	0	192	7	1832	15	0
A10	0	0	0	0	0	0	1	0	5	7287	0
A11	0	0	13	5	0	0	0	0	0	0	13850

(e)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
A1	3072	1187	30	13	8	4	0	0	2	0	0
A2	742	3579	5	13	0	6	0	0	2	3	0
A3	0	8	3428	160	282	21	0	0	0	1	464
A4	94	63	418	3382	279	100	0	0	5	2	11
A5	0	1	605	195	3148	20	0	0	0	2	11
A6	35	21	4	3	1	30739	0	0	0	1	0
A7	0	0	0	0	0	0	30047	21	448	0	0
A8	0	0	0	0	0	0	0	30339	13	0	0
A9	0	0	0	1	0	0	207	12	3845	27	0
A10	0	0	0	0	0	0	0	0	14	14572	0
A11	0	0	215	955	36	0	0	0	0	0	26530

Fig. 7. Confusion matrices for the developed fusion network along with two baseline architectures. The rows (activity label) represent the actual activities, and the columns (predicted label) show the recognized activities. The deeper color presents the better results. (a) CNN ($C2$) (b) ConvLSTM (TS) (c) Fusion ($C1 + TS$) (d) Fusion ($C2 + TS$) (e) Fusion ($C1 + C2 + TS$).

Table 2
Percentile precision (Precision % ± Standard deviation) of the proposed fusion architecture for each activity classification.

Activity	CNN ($C2$)	ConvLSTM (TS)	Fusion ($C1 + TS$)	Fusion ($C2 + TS$)	Fusion ($C1 + C2 + TS$)
$A1$	96.300 ± 0.037	72.800 ± 0.051	86.000 ± 0.062	87.700 ± 0.060	80.299 ± 0.064
$A2$	80.799 ± 0.066	64.700 ± 0.027	74.800 ± 0.058	80.899 ± 0.037	70.100 ± 0.065
$A3$	87.300 ± 0.054	48.200 ± 0.079	74.700 ± 0.042	93.599 ± 0.031	60.400 ± 0.078
$A4$	83.500 ± 0.031	76.800 ± 0.069	86.299 ± 0.029	90.399 ± 0.032	78.600 ± 0.086
$A5$	82.900 ± 0.025	56.700 ± 0.077	82.300 ± 0.027	83.500 ± 0.032	80.500 ± 0.042
$A6$	100.00 ± 0.000	97.100 ± 0.007	99.900 ± 0.003	99.900 ± 0.003	99.600 ± 0.004
$A7$	100.00 ± 0.000	91.799 ± 0.022	98.900 ± 0.003	100.00 ± 0.000	99.100 ± 0.003
$A8$	100.00 ± 0.000	93.499 ± 0.035	99.900 ± 0.003	100.00 ± 0.000	99.800 ± 0.004
$A9$	98.600 ± 0.012	63.199 ± 0.095	90.800 ± 0.045	99.100 ± 0.008	86.300 ± 0.061
$A10$	98.000 ± 0.000	98.400 ± 0.006	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
$A11$	100.00 ± 0.000	95.199 ± 0.009	99.000 ± 0.007	98.000 ± 0.000	98.200 ± 0.008

and fusion of all data achieve 100% recall rate for only two activities ($A8$, and $A10$), and the best recall rate is found from the Fusion ($C2 + TS$) that obtains 100% recall for five activities ($A6$–$A8$, and $A10$–$A11$). Considering the fusion and baseline models, the lowest recall of 59.10% is appraised in ConvLSTM architecture.

Moreover, Table 4 illustrates the F1-Score of the proposed fusion approach along with the baseline models for each activity recognition. Here, the fusion architecture with camera 2 data along with time-series information achieves the best F1-Score (100%) for four activities ($A6$–$A8$, and $A10$) whereas the CNN ($C2$) model obtains 100% F1-Score

Table 3
Percentile recall (Recall % ± Standard deviation) of the proposed fusion architecture for each activity classification.

Activity	CNN ($C2$)	ConvLSTM (TS)	Fusion ($C1+TS$)	Fusion ($C2+TS$)	Fusion ($C1+C2+TS$)
$A1$	78.200 ± 0.091	63.400 ± 0.072	67.699 ± 0.099	79.099 ± 0.063	67.000 ± 0.085
$A2$	85.900 ± 0.082	74.000 ± 0.064	87.500 ± 0.064	89.100 ± 0.054	84.000 ± 0.067
$A3$	88.599 ± 0.025	63.100 ± 0.050	82.999 ± 0.024	80.099 ± 0.043	78.400 ± 0.027
$A4$	78.900 ± 0.044	59.100 ± 0.059	71.600 ± 0.037	83.200 ± 0.033	70.300 ± 0.065
$A5$	94.100 ± 0.025	59.700 ± 0.046	83.700 ± 0.047	96.300 ± 0.022	78.000 ± 0.015
$A6$	100.00 ± 0.000	97.900 ± 0.005	99.900 ± 0.003	100.00 ± 0.000	99.800 ± 0.004
$A7$	100.00 ± 0.000	92.700 ± 0.024	99.869 ± 0.006	100.00 ± 0.000	97.800 ± 0.013
$A8$	100.00 ± 0.000	93.200 ± 0.016	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
$A9$	93.200 ± 0.007	70.500 ± 0.040	90.100 ± 0.017	98.200 ± 0.011	93.100 ± 0.024
$A10$	99.800 ± 0.004	99.200 ± 0.004	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
$A11$	100.00 ± 0.000	87.600 ± 0.010	99.800 ± 0.004	100.00 ± 0.000	93.899 ± 0.011

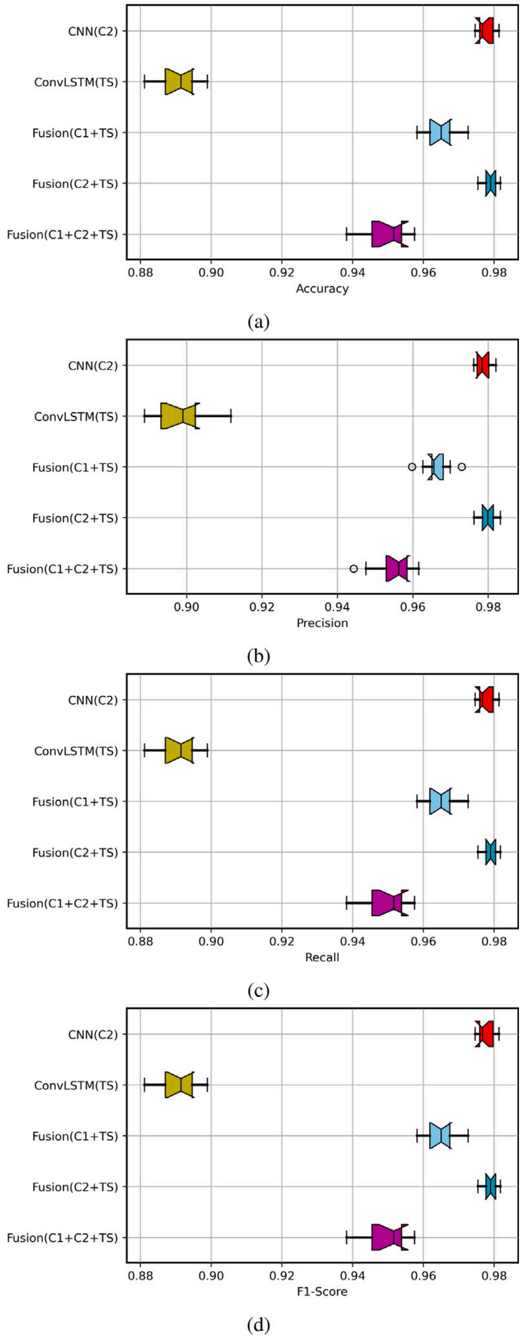


Fig. 8. Boxplot for the comparative analysis of the proposed fusion architecture with the baseline models. (a) Accuracy (b) Precision (c) Recall (d) F1-Score.

for four activates ($A6$–$A8$, and $A11$) but it differs by only one activity. The experimental results reveal that the average F1-Score of 97.90%, 89.10%, 96.50%, 98.00%, and 94.89% are achieved from CNN ($C2$), ConvLSTM (TS), Fusion ($C1 + TS$), Fusion ($C2 + TS$), and Fusion ($C1 + C2 + TS$) architectures, respectively.

Considering the class-wise performances (precision, recall, and F1-Score) of multimodal human activity recognition, it is evident that the performances of human falls are relatively low compared to other ADLs. It could happen due to the nature of the fall events as it consists of standing and laying activities in UP-Fall detection dataset.

Further, the activity recognition results of the test samples for the proposed fusion approach and two baseline models (CNN and ConvLSTM) are shown in Table 5. Two baseline models: CNNs and ConvLSTM are experimented with the visual data from camera 2 and time-series data respectively. The baseline ConvLSTM model achieved a classification accuracy of 89.08%, while the normal CNN obtained relatively better results with a recognition accuracy of 97.76% which almost reaches the recognition level of the fusion architecture. Among the three variants of fusion architecture, the best performances such as accuracy of 96.90%, precision of 97.98%, recall of 96.90%, and F1-Score of 97.88% are obtained from Fusion ($C2 + TS$). While fusing all data, the performances have been somewhat decreased as there are some scenarios with almost the same information. For example, the data from camera 1, and camera 2 contain almost identical activity scenarios but they only differ by view perspectives. The proposed fusion architecture considering the frontal view of the visual data along with time-series information outperforms its corresponding counterparts such as Fusion ($C1 + TS$), and Fusion ($C1 + C2 + TS$) by an accuracy improvement of ∼1.40%, and ∼2.90%, respectively. Moreover, this model suppresses the performance of the baseline architectures (CNN and ConvLSTM) with an accuracy enhancement of ∼0.15%, and ∼8.80%, respectively.

Fig. 8 illustrates the box plots of the performance measures obtained from ten runs of the proposed fusion approach and baseline models for deeper analysis. In all the cases, the CNN ($C2$) and Fusion($C2 + TS$) models outperformed all the other models. In all of the performance metrics, the median values for the Fusion ($C2+TS$) model were better than the CNN ($C2$) model. In all cases, the ConvLSTM (TS) model performed the worst. The Fusion ($C1 + TS$) and Fusion ($C1 + C2 + TS$) models had some outlier values in terms of precision. All the performance measures indicate that the accuracy and reliability are better in the proposed fusion approach with camera 2 and time-series data compared to others.

Fig. 9 presents the Gradient-weighted Class Activation Mapping (Grad-CAM) for each activity of our proposed Fusion ($C2+TS$) model. Grad-CAM is an AI explainability method that visualizes the regions of the input image that the model finds interesting while performing inference. Grad-CAM produces a coarse localization map of the regions that the model deems "important". On all the Grad-CAM of the randomly chosen activity images, the model highlighted the persons. By focusing on the person, their activity can be determined. Thus, the Grad-CAM images indicate that the proposed model is correctly focusing on the correct subject in the images and inferring their activities.

Table 4
Percentile F1-Score (F1-Score % ± Standard deviation) of the proposed fusion architecture for each activity classification.

Activity	CNN ($C2$)	ConvLSTM (TS)	Fusion ($C1+TS$)	Fusion ($C2+TS$)	Fusion ($C1+C2+TS$)
$A1$	85.900 ± 0.046	67.300 ± 0.037	75.000 ± 0.061	82.900 ± 0.021	72.400 ± 0.036
$A2$	82.800 ± 0.043	69.000 ± 0.031	80.399 ± 0.036	84.700 ± 0.023	76.100 ± 0.037
$A3$	87.900 ± 0.033	54.000 ± 0.054	78.499 ± 0.027	86.199 ± 0.024	66.799 ± 0.048
$A4$	80.899 ± 0.024	66.300 ± 0.038	78.100 ± 0.020	86.299 ± 0.016	73.799 ± 0.054
$A5$	88.200 ± 0.010	58.100 ± 0.058	83.000 ± 0.026	89.400 ± 0.018	79.100 ± 0.020
$A6$	100.00 ± 0.000	97.300 ± 0.004	99.900 ± 0.003	100.00 ± 0.000	99.900 ± 0.003
$A7$	100.00 ± 0.000	92.199 ± 0.017	98.700 ± 0.004	100.00 ± 0.000	98.500 ± 0.008
$A8$	100.00 ± 0.000	93.399 ± 0.019	100.00 ± 0.000	100.00 ± 0.000	99.800 ± 0.004
$A9$	95.700 ± 0.009	66.199 ± 0.062	90.299 ± 0.023	98.600 ± 0.004	89.399 ± 0.035
$A10$	98.900 ± 0.003	98.800 ± 0.004	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
$A11$	100.00 ± 0.000	91.200 ± 0.007	99.300 ± 0.004	99.200 ± 0.004	96.000 ± 0.006

Table 5
Activity recognition results of the proposed fusion approach along with baseline models for the test samples.

Architectures	Average				Best			
	Accuracy	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
CNN ($C2$)	97.766 ± 0.002	97.878 ± 0.001	97.766 ± 0.002	97.749 ± 0.002	98.13	98.20	98.13	98.14
ConvLSTM (TS)	89.081 ± 0.005	89.834 ± 0.006	89.081 ± 0.005	89.278 ± 0.005	89.89	91.17	89.89	90.34
Fusion ($C1+TS$)	96.491 ± 0.004	96.625 ± 0.003	96.491 ± 0.004	96.451 ± 0.004	97.26	97.29	97.26	97.22
Fusion ($C2+TS$)	97.902 ± 0.001	97.986 ± 0.002	97.902 ± 0.001	97.880 ± 0.001	98.26	98.30	98.18	98.25
Fusion ($C1+C2+TS$)	94.980 ± 0.006	95.492 ± 0.005	94.980 ± 0.006	95.096 ± 0.006	95.75	96.15	95.75	95.80

4.4. Performance in IoT environment

We evaluated the performance of the models in two IoT deployment scenarios: a smart home environment which represents the same network use case and a virtual care environment which represents the different network use case. The performance in IoT environment was evaluated on the basis of two metrics: latency and response time. Latency and response time depend on various factors including overall network use by other applications, client and server hardware capabilities, as well as the hardware and software configurations of the network. While testing the model performances, we closed all background applications and services.

4.4.1. Smart home environment

Fig. 10 presents the performance of the models in a smart home environment. In this case, the server and the users are on the same network. On average, the latency of the models is 7.5 ms. The response time of the models is 5.5 ms on average. The fusion model trained on image data from camera 2 and time-series data had marginally lower latency and response time compared to the other models. On average, the Fusion ($C2+TS$) model had 0.50 ms lower latency and 5 ms lower response time than the other models.

4.4.2. Virtual care environment

Fig. 11 presents the performance of the models in a virtual care environment. In this scenario, the users and the server are in a different network. The latency of the models is 220 ms on average. The response time of the models is 300 ms on average. All the models, in this case, had similar latency and response time. The latency and response time for the models are higher than the smart home environment results. Latency and response time in different networks are almost always higher than in the same network. This is due to multiple factors such as the distance of the server from the clients, network firewalls, and network configurations. Modern Platform-as-a-Service (PaaS) providers provide multiple services such as servers in multiple global locations, rapid prototyping and deployment, and workload scaling to mitigate some of these issues.

4.5. Comparison with the state-of-the-art

We compare our findings obtained by the fusion architecture with the state-of-the-art to highlight the competitive advantages of our developed approach for multimodal HAR. Table 6 illustrates the performance comparisons of our proposed architecture with the recent studies on UP-Fall detection dataset. We consider the recently published methods possibly 2018 or above as the state-of-the-art techniques. Similar to comparative studies conducted in [32,45], all the performance measures of these state-of-the-art are retrieved from the corresponding articles. The authors in [46] used only the data from multiple wearable sensors and deployed several machine learning algorithms. However, the best recall rate of 69% is obtained using Support Vector Machine (SVM). Only the visual data is utilized in [32,44,45], where the system developed in [44] used the data from camera 1, the camera 2 data is utilized in [32], and the architecture presented in [45] exploited both the lateral and frontal views. It is found that the lowest F1-Score of 71.20% is obtained from [44] using CNN, the best recall rate of 97.95% is achieved in [45] utilizing CNN, and the highest accuracy of 96.70% is appraised in [32] with the use of CNN + GRU. In the state-of-the-art study, only the system introduced in [47] deployed fusion architecture (CNN + LSTM) and used the data from camera 1 as well as wearable sensors. Although the accuracy measure (96.40%) of this framework is good, the other performance measures are not quite remarkable. In our proposed fusion architecture, we used the full set of data of UP-Fall detection dataset. We obtained the best accuracy, precision, and F1-Score of 97.90%, 97.98%, and 97.88%, respectively from the experiments using the frontal view of the camera data, and the full set of wearable sensors information. The recall rate of our developed system is little bit lower than the framework developed in [45].

4.6. Ablation studies

To thoroughly measure the efficiency of our proposed multi-level feature fusion architecture, we conducted several extensive ablation studies considering the different variants of the developed framework. The details description of the models considered for ablation experiments is as follows.

4.6.1. Single-head

In this experiment, we just considered the single head CNN model instead of three-head from our proposed fusion architecture to see the effects of head architectures for feature extraction. The kernel size of this head architecture is 3 × 3.

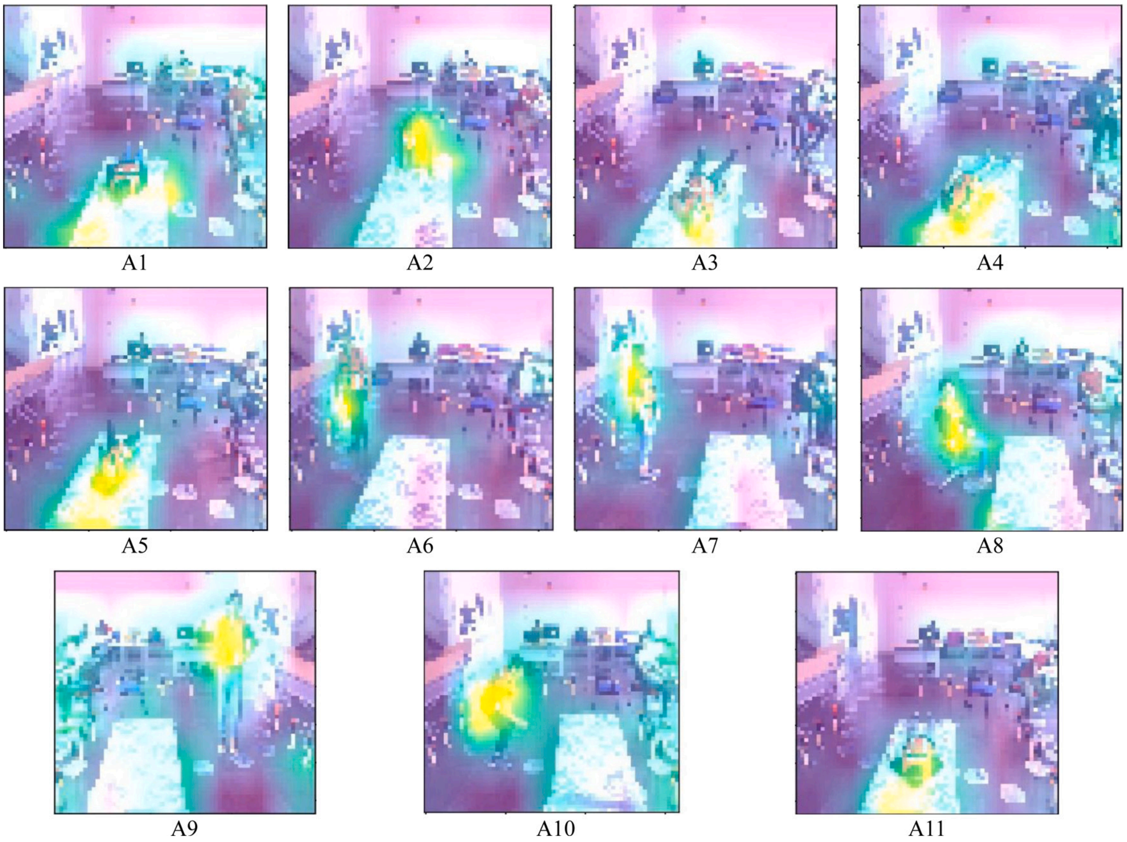


Fig. 9. Grad-CAM of each activity sample on UP-Fall detection dataset for the proposed Fusion ($C2 + TS$) architecture.

Table 6
Performance comparison of the results obtained by the proposed fusion approach and the recent studies on UP-Fall detection dataset.

Refs	Data	Data sources	Architectures	Accuracy	Precision	Recall	F1-Score
[46]	Time-series	Wearable sensors	SVM	–	66%	69%	67%
[44]	Image	Camera 1	CNN	95.10%	71.80%	71.30%	71.20%
[47]	Image + Time-series	Camera 1 and Wearable sensors	CNN + LSTM	96.40%	84.20%	81.50%	82.30%
[45]	Image	Camera 1, 2	CNN	95.64%	96.91%	**97.95%**	97.43%
[32]	Image	Camera 2	CNN + GRU	96.70%	96.90%	96.70%	96.60%
Ours	Image + Time-series	Camera 1 and Wearable sensors	Fusion ($C1 + TS$)	96.49%	96.62%	96.49%	96.45%
		Camera 2 and Wearable sensors	Fusion ($C2 + TS$)	**97.90%**	**97.98%**	97.90%	**97.88%**
		Camera 1, 2 and Wearable sensors	Fusion ($C1 + C2 + TS$)	94.98%	95.49%	94.98%	95.09%

Table 7
Experimental findings of different variants of the proposed fusion architecture on UP-Fall detection dataset.

Architectures	Average				Best			
	Accuracy	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Single-head	74.566 ± 0.046	82.844 ± 0.029	74.566 ± 0.046	75.689 ± 0.0456	81.23	89.03	81.23	83.15
Two-head	97.444 ± 0.003	97.556 ± 0.002	97.444 ± 0.003	97.418 ± 0.003	97.83	97.93	97.83	97.81
No attention	87.841 ± 0.007	89.651 ± 0.005	87.841 ± 0.007	88.346 ± 0.006	88.81	90.55	88.81	89.09
Only channel attention	96.491 ± 0.004	96.625 ± 0.003	96.491 ± 0.004	96.451 ± 0.004	98.18	98.32	98.18	98.17
Only spatial attention	97.358 ± 0.009	97.612 ± 0.006	97.358 ± 0.009	97.358 ± 0.008	98.02	98.08	98.02	97.99

4.6.2. Two-head

In this model, two-head CNN architecture is considered to extract features from the visual data. The kernel size of 3 × 3, and 5 × 5 are considered while omitting the other head architecture with kernel size 7 × 7.

4.6.3. No attention

The proposed fusion architecture is modified in such a way that the CBAM network is omitted from the developed framework. In this ablation experiment, the extracted features from ConvLSTM and three-head CNN architecture are directly fused rather than passing through the CBAM architecture.

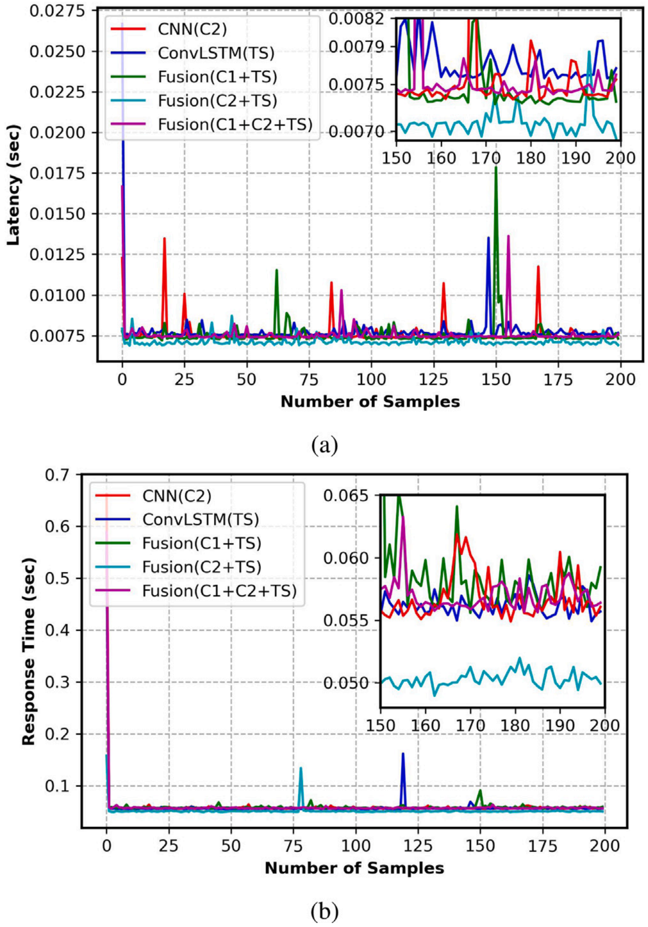
4.6.4. Only Channel Attention

In the proposed fusion architecture, the SAM network is omitted to perform this experiment. The impacts of channel dimension features are evaluated through this experiment.

4.6.5. Only Spatial Attention

The CAM network of the CBAM architecture is deleted from the proposed fusion approach to perform this study. With this experiment, it is realized how the proposed architecture performs without the channel dimension features.

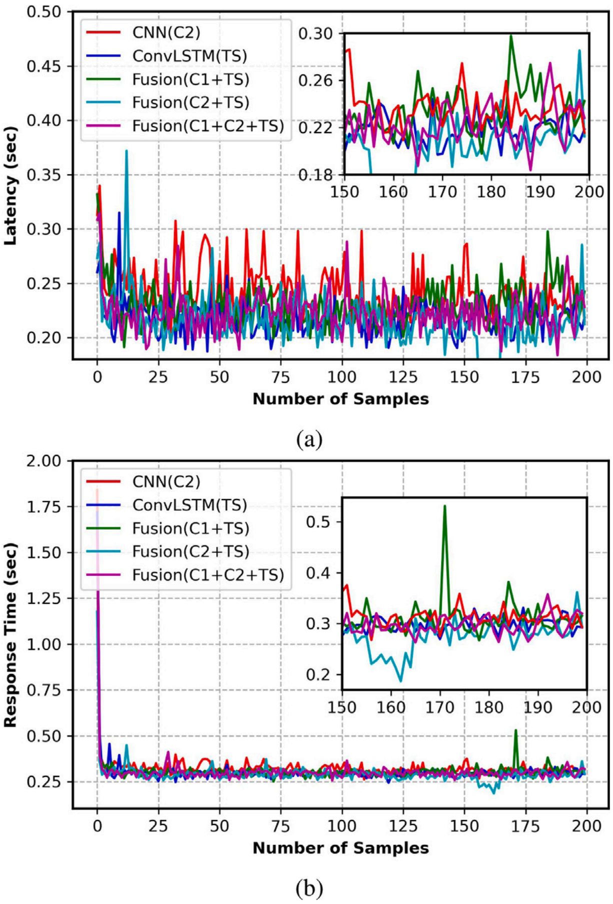
The results obtained by the different variants of the proposed architecture in ablation studies are shown in Table 7. All the experiments in



(a)

(b)

Fig. 10. Performance in smart home environment for multimodal HAR in IoHT. (a) Latency (b) Response time.



(a)

(b)

Fig. 11. Performance in virtual care environment for multimodal HAR in IoHT. (a) Latency (b) Response time.

the alation studies are conducted using the frontal view of the visual data and the time-series data from multiple sensors. Considering the single-head and two-head architecture, the lowest performance (accuracy: 74.56%) is achieved by the single-head architecture compared to the performance (accuracy: 97.44%) of two-head architecture. This has happened as the two-head CNN architecture extracts the relevant features using different kernels, compared to the single-head architecture. The next ablation study is conducted to show the impacts of the attention network on activity recognition performance. The model with no attention block obtained quite low performance approximately ∼10% less accurate than our proposed fusion architecture. The reason for the low performance is that the attention module put more attention on the significant features and removes the unnecessary attributes. Without the use of the attention module, all the features are sent to the classification block through the fusion network; thus decreasing the performance. Lastly, we conducted two other ablation experiments considering the CAM architecture and SAM architecture separately. The fusion architecture with SAM only works slightly better compared to CAM on average. The possible reason that could be mentioned in this scenario is that the SAM network contains several convolution operations while the CAM module consists of a shared MLP network. Consequently, the SAM network retrieved more efficient and high-level features than the CAM architecture.

5. Conclusion

Human activity recognition has become a prominent research area due to the fast development of various emerging computing paradigms and technologies as well as plays a crucial role in IoHT applications by monitoring the physical activities of the individuals. In this article, we proposed a multi-level feature fusion network incorporating the concept of multi-head CNN along with an attention mechanism called CBAM as well as ConvLSTM network for multimodal HAR. This paper enhances the efficiency of multimodal HAR for smart healthcare applications by exploiting ConvLSTM capabilities to model long-term temporal representation from raw multi-sensory information and using multi-head CNN with CBAM to effectively retrieve channel and spatial dimension features from visual information. The extensive experimental findings demonstrated the efficiency and practicability of our proposed fusion architecture compared with two baseline models and state-of-the-art frameworks. The efficiency of the developed framework in IoT platform indicated that this system is quite effective for real-world IoHT applications.

While the proposed architecture performed very well in the two tested deployment scenarios, it still requires huge computational resources to both train and perform inference. Thus, the model is not suitable for edge computing or TinyML use cases. Moreover, the model also requires multimodal data. However, single modality of data is more frequent in the real-world use cases. In the future, more efficient deep learning architecture could lead to new horizon of research capable of handling more complex real-world scenarios. Additionally, this work can be deployed in multiple user scenarios to facilitate scalable healthcare applications.

CRediT authorship contribution statement

Md. Milon Islam: Conceptualization, Methodology, Validation, Investigation, Writing – original draft. **Sheikh Nooruddin:** Methodology, Validation, Investigation, Writing – original draft. **Fakhri Karray:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Ghulam Muhammad:** Investigation, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, Canada. Moreover, G. Muhammad acknowledges the support of Researchers Supporting Project number (RSP2023R34), King Saud University, Riyadh, Saudi Arabia. All authors have read and agreed to this version of the manuscript.

References

[1] Y. Yang, H. Wang, R. Jiang, X. Guo, J. Cheng, Y. Chen, A review of iot-enabled mobile healthcare: Technologies, challenges, and future trends, IEEE Internet Things J. 9 (12) (2022) 9478–9502.

[2] H.F. Nweke, Y.W. Teh, G. Mujtaba, M.A. Al-Garadi, Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions, Inf. Fusion 46 (2019) 147–170.

[3] M. Gjoreski, V. Janko, G. Slapničar, M. Mlakar, N. Reščič, J. Bizjak, V. Drobnič, M. Marinko, N. Mlakar, M. Luštrek, et al., Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors, Inf. Fusion 62 (2020) 47–62.

[4] W. Zheng, L. Yan, C. Gou, F.-Y. Wang, Meta-learning meets the Internet of Things: Graph prototypical models for sensor-based human activity recognition, Inf. Fusion 80 (2022) 1–22.

[5] J. Qi, P. Yang, L. Newcombe, X. Peng, Y. Yang, Z. Zhao, An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure, Inf. Fusion 55 (2020) 269–280.

[6] D. Bouchabou, S.M. Nguyen, C. Lohr, B. LeDuc, I. Kanellos, A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning, Sensors 21 (18) (2021) 6037.

[7] D. Chen, S. Yongchareon, E.M. Lai, Q.Z. Sheng, V. Liesaputra, Locally-weighted ensemble detection-based adaptive random forest classifier for sensor-based online activity recognition for multiple residents, IEEE Internet Things J. 9 (15) (2022) 13077–13085.

[8] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, G. Fortino, Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges, Inf. Fusion 80 (2022) 241–265.

[9] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion 76 (2021) 355–375.

[10] M. Straczkiewicz, P. James, J.-P. Onnela, A systematic review of smartphone-based human activity recognition methods for health research, NPJ Digital Med. 4 (1) (2021) 1–15.

[11] P. Pareek, A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. 54 (3) (2021) 2259–2322.

[12] Y. He, Y. Chen, Y. Hu, B. Zeng, WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi, IEEE Internet Things J. 7 (9) (2020) 8296–8317.

[13] B. Nguyen, Y. Coelho, T. Bastos, S. Krishnan, Trends in human activity recognition with focus on machine learning and power requirements, Mach. Learn. Appl. 5 (2021) 100072.

[14] J. Cao, W. Li, C. Ma, Z. Tao, Optimizing multi-sensor deployment via ensemble pruning for wearable activity recognition, Inf. Fusion 41 (2018) 68–79.

[15] G. Csizmadia, K. Liszkai-Peres, B. Ferdinandy, Á. Miklósi, V. Konok, Human activity recognition of children with wearable devices using lightgbm machine learning, Sci. Rep. 12 (1) (2022) 1–10.

[16] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–20, http://dx.doi.org/10.1109/TPAMI.2022.3183112.

[17] Y. Wu, H. Cao, G. Yang, T. Lu, S. Wan, Digital twin of intelligent small surface defect detection with cyber-manufacturing systems, ACM Trans. Internet Technol. (2022) 1–20.

[18] M.M. Islam, S. Nooruddin, F. Karray, G. Muhammad, Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects, Comput. Biol. Med. 149 (2022) 106060.

[19] L. Werthen-Brabants, G. Bhavanasi, I. Couckuyt, T. Dhaene, D. Deschrijver, Split BiRNN for real-time activity recognition using radar and deep learning, Sci. Rep. 12 (1) (2022) 1–11.

[20] O. Nafea, W. Abdul, G. Muhammad, M. Alsulaiman, Sensor-based human activity recognition with spatio-temporal deep learning, Sensors 21 (6) (2021) 2141.

[21] C.-L. Yang, Z.-X. Chen, C.-Y. Yang, Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images, Sensors 20 (1) (2019) 168.

[22] A. de Santana Correia, E.L. Colombini, Attention, please! A survey of neural attention models in deep learning, Artif. Intell. Rev. 55 (8) (2022) 6037–6124.

[23] E. Garcia-Ceja, C.E. Galván-Tejada, R. Brena, Multi-view stacking for activity recognition with sound and accelerometer data, Inf. Fusion 40 (2018) 45–56.

[24] Q. Li, R. Gravina, Y. Li, S.H. Alsamhi, F. Sun, G. Fortino, Multi-user activity recognition: Challenges and opportunities, Inf. Fusion 63 (2020) 121–135.

[25] M.A. Al-qaness, A. Dahou, M. Abd Elaziz, A. Helmi, Multi-ResAtt: Multilevel residual network with attention for human activity recognition using wearable sensors, IEEE Trans. Ind. Inform. 19 (1) (2023) 144–152.

[26] J. Lu, X. Zheng, M. Sheng, J. Jin, S. Yu, Efficient human activity recognition using a single wearable sensor, IEEE Internet Things J. 7 (11) (2020) 11137–11146.

[27] X. Zhou, W. Liang, I. Kevin, K. Wang, H. Wang, L.T. Yang, Q. Jin, Deep-learning-enhanced human activity recognition for internet of healthcare things, IEEE Internet Things J. 7 (7) (2020) 6429–6438.

[28] M. Abdel-Basset, H. Hawash, V. Chang, R.K. Chakrabortty, M. Ryan, Deep learning for heterogeneous human activity recognition in complex iot applications, IEEE Internet Things J. 9 (8) (2022) 5653–5665.

[29] M. Abdel-Basset, H. Hawash, R.K. Chakrabortty, M. Ryan, M. Elhoseny, H. Song, ST-DeepHAR: Deep learning model for human activity recognition in ioht applications, IEEE Internet Things J. 8 (6) (2020) 4969–4979.

[30] H. Zhang, Z. Xiao, J. Wang, F. Li, E. Szczerbicki, A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention, IEEE Internet Things J. 7 (2) (2019) 1072–1080.

[31] S.K. Yadav, K. Tiwari, H.M. Pandey, S.A. Akbar, A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions, Knowl.-Based Syst. 223 (2021) 106970.

[32] S.K. Yadav, A. Luthra, K. Tiwari, H.M. Pandey, S.A. Akbar, ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling, Knowl.-Based Syst. 239 (2022) 107948.

[33] H. Ramirez, S.A. Velastin, I. Meza, E. Fabregas, D. Makris, G. Farias, Fall detection and activity recognition using human skeleton features, IEEE Access 9 (2021) 33532–33542.

[34] A.R. Inturi, V. Manikandan, V. Garrapally, A novel vision-based fall detection scheme using keypoints of human skeleton with long short-term memory network, Arab. J. Sci. Eng. (2022) 1–13.

[35] F. Lin, Z. Wang, H. yu Zhao, S. Qiu, X. Shi, L. Wu, G. Fortino, R. Gravina, Adaptive multimodal fusion framework for activity monitoring of people with mobility disability, IEEE J. Biomed. Health Inf. 26 (8) (2022) 4314–4324.

[36] C.M. Ranieri, P.A. Vargas, R.A. Romero, Uncovering human multimodal activity recognition with a deep learning approach, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–8.

[37] C.M. Ranieri, S. MacLeod, M. Dragone, P.A. Vargas, R.A.F. Romero, Activity recognition for ambient assisted living with videos, inertial units and ambient sensors, Sensors 21 (3) (2021) 768.

[38] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, K.-K.R. Choo, Adaptive fusion and category-level dictionary learning model for multiview human action recognition, IEEE Internet Things J. 6 (6) (2019) 9280–9293.

[39] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 33 (12) (2022) 6999–7019.

[40] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.

[41] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. 29 (9) (2017) 2352–2449.

[42] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[43] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, W.-c. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS, 2015, pp. 802–810.

[44] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, C. Peñafort-Asturiano, UP-fall detection dataset: A multimodal approach, Sensors 19 (9) (2019) 1988.

[45] R. Espinosa, H. Ponce, S. Gutiérrez, L. Martínez-Villaseñor, J. Brieva, E. Moya-Albor, A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-fall detection dataset, Comput. Biol. Med. 115 (2019) 103520.

[46] L. Martínez-Villaseñor, H. Ponce, R.A. Espinosa-Loera, Multimodal database for human activity recognition and fall detection, Multidiscip. Digital Publ. Inst. Proc. 2 (19) (2018) 1237.

[47] L. Martínez-Villaseñor, H. Ponce, K. Perez-Daniel, Deep learning for multimodal fall detection, in: 2019 IEEE International Conference on Systems, Man and Cybernetics, SMC, IEEE, 2019, pp. 3422–3429.