ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/366180314

GRU-INC: An inception-attention based approach using GRU for human activity recognition

Article *in* Expert Systems with Applications · December 2022

DOI: 10.1016/j.eswa.2022.119419

CITATIONS
66

READS
185

8 authors, including:

Taima Rahman Mim

2 PUBLICATIONS **69** CITATIONS

SEE PROFILE

Sadia Afreen
Boise State University

3 PUBLICATIONS **166** CITATIONS

SEE PROFILE

Mohammad Abu Yousuf
Jahangirnagar University

130 PUBLICATIONS **2,071** CITATIONS

SEE PROFILE

Shahadat Uddin
The University of Sydney

181 PUBLICATIONS **5,905** CITATIONS

SEE PROFILE

All content following this page was uploaded by Mohammad Abu Yousuf on 18 July 2024.

The user has requested enhancement of the downloaded file.

Highlights

GRU-INC: An Inception-Attention based Approach using GRU for Human Activity Recognition

Taima Rahman Mim,Maliha Amatullah,Sadia Afreen,Mohammad Abu Yousuf,Shahadat Uddin,Salem A. Alyami,Khondokar Fida Hasan,Mohammad Ali Moni

- A GRU-Inception based deep learning model to identify human activities
- Attention mechanism is incorporated with GRU to improve temporal feature extraction
- Inception modules along with a CBAM block further highlights the spatial features
- The model is relatively wider rather than a deep structure thus reducing complexity

GRU-INC: An Inception-Attention based Approach using GRU for Human Activity Recognition

Taima Rahman Mim[a], Maliha Amatullah[a], Sadia Afreen[a], Mohammad Abu Yousuf[b,*], Shahadat Uddin[c], Salem A. Alyami[d], Khondokar Fida Hasan[e] and Mohammad Ali Moni[f,**]

[a] *Information and Communication Technology, Bangladesh University Of Professionals, Mirpur Cantonment, Dhaka, Bangladesh*

[b] *Institute of Information Technology, Jahangirnagar University, Savar, Dhaka, Bangladesh*

[c] *Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Darlington,NSW, 2008,, Australia*

[d] *Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 13318, Saudi Arabia*

[e] *School of Computer Science, Queensland University of Technology (QUT), 2 George Street, Brisbane 4000, Australia, Australia*

[f] *Artificial Intelligence & Data Science, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland St Lucia, QLD 4072, Australia*

ARTICLE INFO

Keywords:
Human Activity Recognition (HAR)
Inception Module
Convolutional Block Attention Module (CBAM)
Gated Recurrent Unit (GRU)
Attention Mechanism

ABSTRACT

Human Activity Recognition (HAR) is very useful for the clinical applications, and many machine learning algorithms have been successfully implemented to achieve high-performance results. Although handcrafted feature extraction techniques were used in the past, Artificial Neural Network (ANN) is now more popular. In this work, a model has been proposed called Gated Recurrent Unit-Inception (GRU-INC) model has been proposed, which is an Inception-Attention based approach using Gated Recurrent Unit (GRU) that effectively makes use of the temporal and spatial information of the time-series data. The proposed model achieved an F1-score of 96.27%, 90.05%, 90.30%, 99.12%, and 95.99% on the publicly available datasets such as, UCI-HAR, OPPORTUNITY, PAMAP2, WISDM, and Daphnet, respectively. GRU along with Attention Mechanism (AM) was utilized for the temporal part, and Inception module along with Convolutional Block Attention Module (CBAM) was exploited for the spatial part of the model. The proposed architecture was evaluated against state-of-the-art models and similar works. It has been proved that the GRU-INC model has a higher recognition rate as well as lower computational cost. Thus our framework could be applicable in activity associated clinical and rehabilitation applications.

1. Introduction

Human activity recognition is known as the problem of identifying various kinds of daily human activity collected either via image and video sequences or various types of sensors. It has become one of the most popular scopes of study due to its requirement to identify human activity for healthcare, security, smart home, surveillance, sports, etc. One of the most crucial needs of HAR is detecting sudden changes in the activity of elderly/sick people to take immediate action. Various techniques have been used to date to achieve more reliable and precise results. For human activity recognition, both image-based and sensor-based data are used. Data collected from wearable and smartphone embedded sensors like accelerometer, gyroscope, magnetometer, Global Positioning System (GPS), etc. represent time series data. In detecting activity, low-level data are collected from sensors that undergo a series of pre-processing techniques to obtain high-level information about those activities. Sensor-based HAR is more popular due to the fact that sensor data is lightweight in nature, and the collection of sensor data is faster, precise, and privacy is less hindered compared to that of image-based data. Initially, handcrafted methods like K-Nearest Neighbor (KNN) (Ehatisham-Ul-Haq, Javed, Azam, Malik, Irtaza, Lee & Mahmood, 2019), Support Vector Machine (SVM), Naïve Bayes (Ghazal,

*Corresponding author

**Principal corresponding author

taimarh@gmail.com (T.R. Mim); amatullahmaliha15@gmail.com (M. Amatullah); s.afreen07@gmail.com (S. Afreen); yousuf@juniv.edu (M.A. Yousuf); shahadat.uddin@sydney.edu.au (S. Uddin); saalyami@imamu.edu.sa (S.A. Alyami); fida.hasan@qut.edu.au (K.F. Hasan); m.moni@uq.edu.au (M.A. Moni)

ORCID(s): 0000-0003-0101-4749 (M.A. Yousuf); 0000-0003-0091-6919 (S. Uddin); 0000-0002-5507-9399 (S.A. Alyami); 0000-0002-8008-8203 (K.F. Hasan); 0000-0003-0756-1006 (M.A. Moni)

Khan, Mubasher Saleem, Rashid & Iqbal, 2019), Principal Component Analysis (PCA) (Hassan, Uddin, Mohamed & Almogren, 2018) etc. methods were used for the feature extraction purpose.

Since the handcrafted feature extraction methods require human effort and are more time-consuming, neural networks were introduced. Artificial Neural Networks possess the capability of learning through examples and events and producing outputs based on that learning; they can even produce accurate output without having complete knowledge. The efficiency and performance of a neural network increase when the number of neurons is increased, but this might also hinder the generalization capability and cause the overfitting problem. Examples of neural networks are Convolutional Neural Network (CNN), variants of CNN such as Inception, Residual Neural Network ResNet etc., and Recurrent Neural Networks (RNN) such as GRU, Long Short-Term Memory (LSTM).

CNN was initially used only in image detection (Bevilacqua, MacDonald, Rangarej, Widjaya, Caulfield & Kechadi, 2019) but Münzner, Schmidt, Reiss, Hanselmann, Stiefelhagen & Dürichen (2017) achieved significant results as a result of implementing CNN on sensor data. Several researches have been done by using CNN on sensor data as well as integrating CNN with RNNs (Xia, Huang & Wang, 2020), (Shojaedini & Beirami, 2020) using variants of CNN like ResNet in (Zhang, Qiao, Lin & Zhou, 2021), Inception module in (Xu, Chai, He, Zhang & Duan, 2019) and both of these variants in (Ronald, Poulose & Han, 2021). It is highly important to secure the temporal dependencies of time-series data. Consequently, recurrent neural networks are proved to be capable of such a purpose. LSTM and GRU are the two kinds of RNNs that are widely used. GRU has a lesser number of gates than LSTM and uses fewer parameters, as a result of which GRU is more popular now. GRU is used alongside CNN or any of its variants (Xu et al., 2019) and showed great performance in activity detection. Although the attention mechanism was initially used for Natural Language Processing (NLP), later, it was used in various fields, including HAR. Again to secure the temporal relationship as well as the long-term dependencies between data, the attention mechanism incorporated with GRU (Haque, Tonmoy, Mahmud, Ali, Khan & Shoyaib, 2019), (Ma, Li, Zhang, Gao & Lu, 2019), (Zhang, Xiao, Wang, Li & Szczerbicki, 2019) proved to be exceedingly successful. Woo, Park, Lee & Kweon (2018) introduced CBAM for the first time where they deduced channel and spatial attention maps from a given feature map with the purpose of more precisely refining the feature extraction technique. In (Gao, Zhang, Teng, He & Wu, 2021) an amalgamation of the channel and temporal attention caused the enhancement of CNN, as well as taking the advantage of ResNet for better extraction of features.

Human activity recognition is a significant part of daily life due to its major contribution in various fields such as: healthcare (Taylor, Shah, Dashtipour, Zahid, Abbasi & Imran, 2020): in order to detect a sudden change in the activity of a patient in hospitals, smart home (Mekruksavanich & Jitpattanakul, 2021): which allows the smart devices to automate as per the human mind based on its predictions, surveillance of elderly and sick people (Schrader, Toro, Konietzny, Rüping, Schäpers, Steinböck, Krewer, Mueller, Guettler & Bock, 2020), Internet of Things (IoT) (Zhou, Liang, Wang, Wang, Yang & Jin, 2020): in order to provide personalized support to specific individuals, detection of aggressive activity in prison (Zhang, Xu, Xiong, Sun, Shi, Fan & Li, 2020). Existing approaches are not fully effective in detecting the complex activities and sudden transition of state. As a result, real-time application of human activity recognition has become mandatory and new technologies need to be implemented in this field. The research questions that are addressed in this work are summarized below:

1. How can a Deep Learning (DL) model be created for human activity recognition?
2. How to improve the feature extraction for multi-modal sensor-based data?
3. How to verify the model performance?

The paper introduces a deep learning-based model which is both efficient and more precise in recognizing complex activities. The proposed GRU-INC is a combination of GRU and Inception, respectively, for the temporal and spatial parts. GRU, along with the Attention Mechanism that has been used in the temporal part, can highly improve feature extraction. The Inception module with an addition of a CBAM block has been used in the spatial part, which further highlights the feature map. To achieve higher efficiency, the proposed model uses a Global Average Pooling (GAP) layer that comparatively decreases the number of model parameters and enables the model to converge faster than the existing models. The proposed model has been evaluated against five publicly available datasets: UCI-HAR, OPPORTUNITY, PAMAP2, WISDM, and Daphnet.

This paper's primary contributions can be summarized in the following points:

1. The proposed GRU-INC model is an efficient DL approach consisting Gated Recurrent Unit-Attention Mechanism (GRU-AM) in the temporal part and Inception-Convolutional Block Attention Module (Inception-CBAM) in the spatial part that achieved remarkable performance lessening the complexity of the model.

2. CBAM has been implemented along with Inception module in time-series data for the first time in human activity recognition.
3. The proposed model is proven to be generalized, being evaluated on five publicly available datasets and achieving higher accuracy and minimal loss.
4. By conducting extensive experiments, it is validated that GRU-INC is an effective model that is comparatively lesser deep, rather wide. It shows better performance due to the reduction of complexity and is justified by the comparison made with the existing models and the state-of-the-art.

This paper has been divided into the following sections. Section II comprises the background study necessary for this research and a summary of previous research on human activity recognition. Section III explains the proposed model by giving a thorough description of the pre-processing, the model architecture, and the feature extraction of both the spatial and temporal parts. The results have been evaluated with the datasets and existing models in Section IV. Finally, Section V contains the conclusion of the research work, its limitations, and works that can be extended in the future.

2. Related Work

Different techniques and models have been implemented by various researchers in order to achieve better results on the recognition rate, although higher accuracy does not always interpret that the model is efficient. By compromising the complexity and efficiency of the model, a better accuracy rate might be achieved, but that is not feasible. As a result, it has become a challenge to create a model that is efficient in terms of computation and produces good results based on accuracy.

The most challenging part of human activity recognition is the variation of performing the same activity done by different individuals. Ehatisham-Ul-Haq et al. (2019) proposed a viable multimodal feature-level fusion approach for HAR using wearable inertial sensors, RGB, and depth camera sensors for classifying 27 human actions using KNN classifier. They achieved the overall best performance with an accuracy rate of 97.6%, which is better than the state of the art. They Hassan et al. (2018) presented a smartphone inertial sensors-based approach for human activity recognition. From the sensor signals, multiple robust features have been extracted followed by Kernel Principal Component Analysis (KPCA) for dimension reduction and then n combined with deep learning technique, Deep Belief Network (DBN) for activity training and recognition. A deep learning-based approach was applied, Alema Khatun & Abu Yousuf (2020) which found that the CNN model performs comparatively better in HAR than random forest and support vector machine algorithms. The CNN model here achieved 99% accuracy, which is 12% and 26% more than SVM and random forest, respectively. In order to make use of the "personal" nature of activities, Buffelli & Vandin (2021) came up with a transfer learning technique to create a model that can be adapted for an individual user and detect similar kinds of activities like walking up and downstairs. Hernández, Suárez, Villamizar & Altuve (2019) introduced the use of bidirectional RNN; however, it produces an overall accuracy of 92.67%, but it is able to differentiate between similar activities like walking upstairs and downstairs. Shojaedini & Beirami (2020) proposed an algorithm that minimized the accuracy saturation phenomenon along with improving the optimization ability of LSTM-CNN. The accuracy of the classification of several activities was increased by using the proposed structure, achieving 96.32% and 82.38% accuracy on two sets of WISDM. Hybrid models have been used in order to detect activity, and such a model proposed Abbaspour, Fotouhi, Sedaghatbaf, Fotouhi, Vahabi & Linden (2020) achieved high accuracy of about 99.57%, 99.6%, 99.65%, and 99.8%, respectively, on the publicly available dataset PAMAP2 on four kinds of hybrid models, although it compromised with the training. The reason for achieving high accuracy was due to the use of bi-directional RNNs; at the same time, this caused an increase in computational time as both past to future and future to past data are being processed. A CNN model integrated with Bi-directional Long Short-Term Memory (BiLSTM) was proposed Nafea, Abdul, Muhammad & Alsulaiman (2021). It extracted features from sensor data as well as effectively selected optimal video representation. The model achieved high precision with an accuracy of 98.53% in the WISDM dataset.

Gao et al. (2021) introduced DanHAR, an attention-based residual network for both blending channel and temporal attention module, which improved the representation power of CNN by blending channel and temporal attention. They achieved the best accuracy of 98.85% for the WISDM dataset. DeepSense was introduced in 2017, which integrated CNN and RNN. It found that interactions among sensing modalities and different modalities could be used to train a model in order to extract robust features. However, it required very high human effort (Yao, Hu, Zhao, Zhang & Abdelzaher, 2017). Using the motivation behind DeepSense and Transformer introduced by Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin (2017), TRASEND was proposed, (Buffelli & Vandin, 2021) which used a pure attention mechanism without the use of a recurrent network and achieved an overall accuracy of 84.8%, 72.3%, and 70.2% on publicly available datasets H-HAR, PAMAP2 and USC-HAD by validating the technique of data augmentation for increasing the generalization capability of the model. The use of attention mechanisms in human activity recognition for sensor data has shown significant results. The attention mechanism is not only used in RNN but also along with both CNN and RNN in a model called AttnSense proposed by Ma et al. (2019) where they achieved significantly good results by capturing both spatial and temporal dependencies as precisely capturing timesteps. Zhang et al. (2020) proposed a new architecture SMFE-SCbSE which inherently used attention mechanism as well as squeeze and excitation methods to re-calibrate features into both space and channel and achieved impressive results. On the other hand, by explicitly using the attention mechanism and squeeze and excite methods along with residual networks for capturing spatial features, ST-DeepHAR was introduced, (Abdel-Basset, Hawash, Chakrabortty, Ryan, Elhoseny & Song, 2020) which, without using any pre-processing methods except data normalization, was able to effectively infer human activity by showing great computational efficiency achieving exceptionally high accuracy on the UCI-HAR and WISDM datasets. A fast and robust computational model was proposed, (Qi, Su, Yang, Ferrigno, De Momi & Aliverti, 2019) which was created to integrate signal processing algorithms and signal selection modules as well as to add a data compression module that proved its efficiency in terms of faster computation and an accuracy of 94.18%.

In (Zhang et al., 2019), although they achieved a high recognition rate of 96.4% due to using a multi-head convolutional network, their computational complexity was higher. Inception network has not been explored much in human activity recognition using sensor data. By integrating Inception network along with the Residual network Ronald et al. (2021) proposed a unique method that showed remarkable performance when compared with the existing deep learning methods. This model also achieved higher accuracy using lesser device resources. The use of Inception Neural Network in InnoHAR developed by Xu et al. (2019) showed an accuracy of 94.6%, 93.5%, and 94.5% but a slight declination in the prediction results, which somehow occurred due to the deep structure of the network. However, Zhang et al. (2021) used Res-Net, where the network layers were limited to a specific range to prevent overfitting and model performance degradation. This lessened the computational complexity, but recognition rate was not that high since the learned features were not very deep. Xia et al. (2020) applied a unique technique by using a GAP layer replacing the Fully Connected (FC) layer. Thus resulted in reducing the model parameters at the same time, ensuring a good recognition rate of about 95.78%, 95.85%, and 92.63%.

In this paper, we focused on achieving satisfactory results based on accuracy and, at the same time, reducing the computational complexity and training time.

Table 1: Recent works on Human activity recognition based on diverse methods

Method	**Ref.**	**Main Contribution**	**Limitation**
Handcrafted	Ehatisham-Ul-Haq et al. (2019)	Proposed a KNN classifier with multimodal feature-level fusion of dense Histogram of Oriented Gradients (HOG) features from RGB & depth video data with time-domain features from inertial sensor data achieving an accuracy rate of 97.6%.	Used non-practical pre-segmented actions and lack of specific application of the proposed fusion framework.
	Hassan et al. (2018)	Presented a smartphone inertial sensors-based approach, followed by KPCA for dimension reduction and then combined with Deep Belief Network (DBN) for activity training and recognition which resulted in an accuracy of 95.85%.	Less efficient and less complex activity's recognition in real-time environments.
CNN-RNN based	Xia et al. (2020)	Proposed a Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) model with high robustness and better activity detection capability with fewer parameters. The model was best evaluated on WISDM and achieved an overall accuracy of 95.85%.	Unadaptable to extract activity features. Reduced computational speed due to the LSTM layers.

Continued on next page

Table 1 – *Continued from previous page*

Method	Ref.	Main Contribution	Limitation
	Shojaedini & Beirami (2020)	Proposed an algorithm that minimizes the accuracy saturation phenomenon along with improving the optimization ability of LSTM-CNN, achieving 96.32% and 82.38% accuracy on two sets of WISDM.	Showed a slight decrement in accuracy for non-challenging activities.
Inception & ResNet	Ronald et al. (2021)	Proposed iSPLInception, which is a Inception-ResNet based model and achieved the highest accuracy for UCI-HAR, which is 95.09%.	Performance was notably decreased due to a large number of parameters.
	Xu et al. (2019)	Proposed an InnoHAR model based on the combination of Inception Neural Network and GRU by concatenating convolution kernels of different scales and splicing with max-pooling layers. The best accuracy achieved was 94.6% for OPPOTUNITY dataset.	Adjusting the network structure, including the size of kernels and the connection method. Imbalanced dataset.
Attention Mechanism	Gao et al. (2021)	Proposed an attention-based residual network for both blending channel and temporal attention module with the best accuracy of 98.85% for WISDM dataset.	Higher computational complexity due to the huge number of parameters.
	Zhang et al. (2020)	Proposed a multi-feature extraction framework using an attention mechanism to increase accuracy and GRU and the squeeze-and-excite blocks helped to distinguish similar activities. Achieved 99.10% accuracy on AAD dataset.	Did not cover all types of activities.
	Ma et al. (2019)	Proposed a fusion of attention mechanism with CNN and RNN to capture both temporal and spatial dependencies. AttnSense performed best for Heterogeneous dataset with 96.5% accuracy.	Though the performance resulted to be good, the complexity rose at the expense of high parameters.

3. Methodology

3.1. Pre-Processing

Raw signal data compiled from motion sensors such as body-worn, Inertial Measurement Units (IMU), smartphone sensors, etc. are in 2D format, which may include several ineffective and diverse data dimensions. Pre-Processing is necessary to refine the data and prepare it to feed into the model. Here pre-processing has been done in the following four steps: 1. Linear Interpolation to recover the missing values, 2. Z-score standardization for regulating the values on the same scale, 3. Data Augmentation to increase the size of datasets, and finally, 4. Segmentation to split the input sequences into fixed-length segments. After the pre-processing stage, the raw 2D signal data is ready to feed into the model in 3D segments.

3.1.1. Linear Interpolation

Since the data is collected with the help of embedded and body-worn sensors, there might be a chance of losing a few data. In order to recover the lost data, linear interpolation method is used, which uses an algorithm to fill up the missing values by predicting based on the present values.

3.1.2. Z-score standardization

The z-score standardization is done in order to standardize the values on the same scale. It is done by dividing the deviation of a score by the standard deviation of that dataset. As a result, the mean becomes 0, and the standard deviation becomes 1. That is, the values are centered around the mean having a standard deviation of 1. The equation

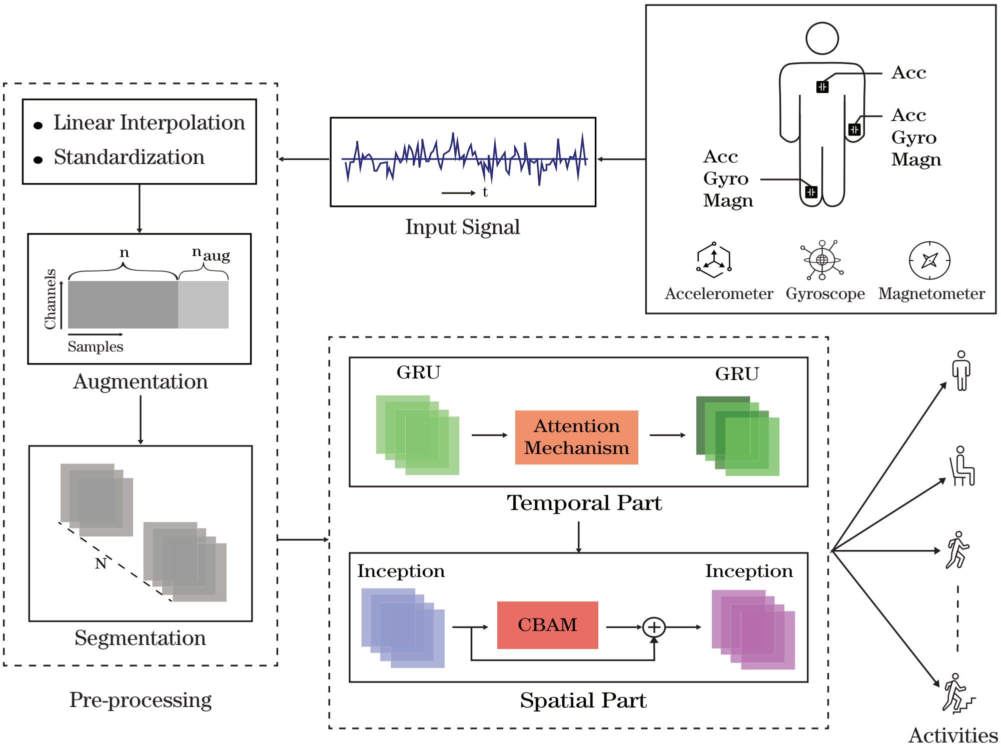


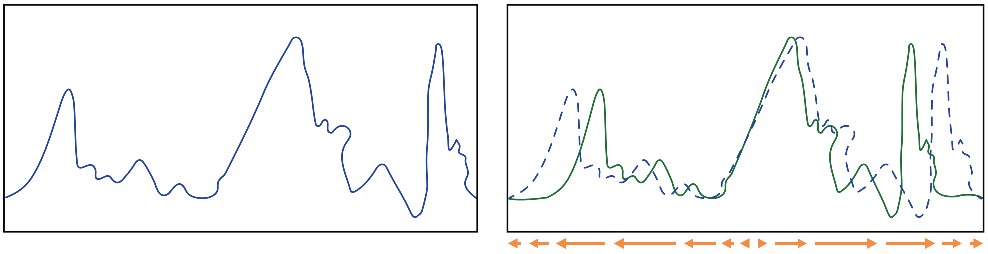
Figure 1: Workflow of the proposed model

for Z-score standardization is given in the equation (1).

$$z(ij) = \frac{a(ij) - \mu}{a} \tag{1}$$

3.1.3. Data Augmentation

Data augmentation is performed in order to increase the training data size, thus advancing a model's learning capability and improving model performance. One of the methods of data augmentation is Time-Warping. It is a way of perturbing the temporal location of samples within the window. Um, Pfister, Pichler, Endo, Lang, Hirche, Fietzek & Kulić (2017) achieved augmentation by altering the time intervals in a smooth manner, as a result changing the temporal locations of the samples. The process of data augmentation is depicted in Fig 2.



a) Original

b) Time Warping

Figure 2: Data augmentation using time warping

3.1.4. Segmentation

For time-series data, segmentation is a technique of splitting the input sequence into discrete slices with the purpose of preserving the temporal relationship between the data points. Different lengths of the sliding windows as well as different overlap values were used for different datasets, and the most optimal length was fixed.

3.2. Model Architecture

In the proposed approach, GRU-INC: An Inception-Attention based approach using GRU for Human Activity Recognition is put forward in which the model architecture is composed of two parts: i) Temporal feature extraction and ii) Spatial feature extraction. Fig 1 illustrates the total process of the GRU-INC model.

3.2.1. Temporal Feature Extraction

RNN is capable of taking the benefit of the temporal relationship between sensor data in a time sequence, Xia et al. (2020) and since during feature extraction, RNN depends on both the past and present data, RNN is utilized specifically for its efficacy with time-series data. In the temporal feature extraction layer, GRU is chosen as the suitable RNN for the proposed work. GRU has a special advantage over LSTM in terms of the number of gates Chung, Gulcehre, Cho & Bengio (2014). GRU uses lesser parameters during training, as a result of which it utilizes lesser memory, and execution time is faster.

To successfully extract the temporal features, two GRU blocks are used, including an Attention block between them. The numbers of neurons of the GRU blocks are set to 256 and 244, respectively. The pre-processed data is fed into the first GRU block.

GRU obtains different modalities of the pre-processed data and adjusts the information flow by using two gates (update gate and reset gate) that determine the information supplied to the output. The reset gate r_t regulates how the signal input x_t is combined with the prior hidden state h_{t-1} (2), and the update gate z_t determines how much of the previous memory, h_{t-1} is preserved (3). With the help of GRU, the relevant data is retained to output one hidden state, h, for each input time step and therefore transmitted to the subsequent network time steps.

$$z_t = \sigma(x_t \cdot W^z + h_{t-1} \cdot U^z) \tag{2}$$

$$r_t = \sigma(x_t \cdot W^r + h_{t-1} \cdot U^r) \tag{3}$$

$$h'_t = tanh(x_t \cdot W^z + (r_t * h_{t-1})U^z) \tag{4}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t \cdot h'_t \tag{5}$$

Here * denotes element-wise multiplication.

Attention mechanism incorporated with the GRU block emphasizes the output features obtained and impacts more during classification. The attention module pays more “attention” to the important modalities and highlights the important features of the GRU (Ma et al., 2019).

Here, the output hidden states h_t of GRU are taken as input, and a random weight (W) is generated. The addition of the product of h_t and W with the bias, followed by a tanh operation on the total, yields an attention weight (e) for each modality (6). To scale down e, it is passed through a softmax function to get a normalized weight, a (7). The input data utilizes multiple sensors, and in order to detect movements accurately, attention weights are employed to reflect the relative relevance of different sensors. The output feature map is produced by the multiplication of the normalized attention weights with the original input, h_t (8).

$$e = tanh(h_t \cdot W + b) \tag{6}$$

$$a = softmax(e) \tag{7}$$

$$Output = h_t * a \tag{8}$$

After the attention mechanism, the output is fed into another block of GRU to allow for greater complexity. As the input is already the result of a GRU layer, the current GRU creates a complex feature representation of the current input. The second GRU can detect reasonably complex features by the weight differences between the important modalities and the inconsequential ones from the attention-fused block.

3.2.2. Spatial Feature Extraction

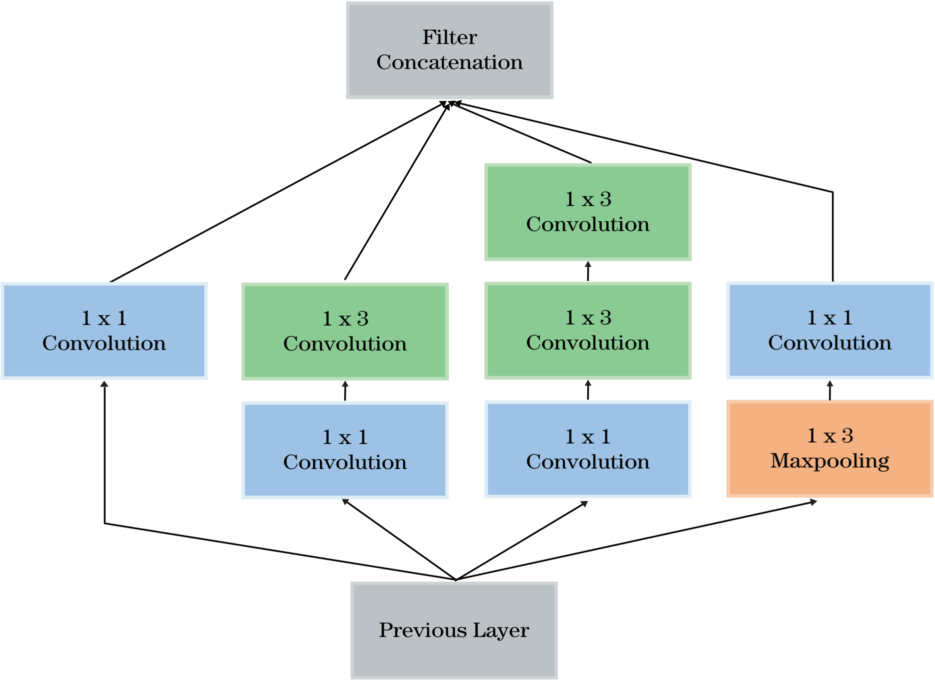


Figure 3: Block diagram of Inception V2

In the spatial part, the feature map obtained from the temporal part is used as the input of the two consecutive Inception blocks followed by a MaxPool1D layer. The proposed method uses Inception Network for the purpose of making the network deep as well as wide in a structure, thus acquiring a minimal increase in computation load. Incorporation of Attention with Inception network is implemented in this model, which is unprecedented. Among the three versions of Inception, Inception v2 module is chosen with the aim to carry out the computations as efficiently as possible with the help of factorized convolutions proposed by Szegedy, Vanhoucke, Ioffe, Shlens & Wojna (2016). In the modified Inception architecture, Fig 3, one-dimensional convolution (Conv1D) has been used to convolve with the temporal dimension of the sensor waveform data. Conv1D is highly resistant to noise and is particularly useful for recovering the latent properties of signals shaped in a single dimension, such as the input time series data. Consequently, it is chosen as a component of the proposed model for extracting features from sensor data of varying lengths. Components of the modified Inception architecture, Fig 3, includes 1x1, 1x3, and 1x5 convolutional layers, max-pooling layer, and concatenation layer. The 1x1 convolutional layer reduces the dimension of the feature map and learns patterns across the dimensions. Here, one 3x3 and one 5x5 convolution kernel of the original Inception v2 were replaced by three 1x3 convolution kernels where two of them were cascaded parallelly, replacing the 5x5. Two consecutive 1x3 perform better than one 1x5 convolution by decreasing the computational time as well as increasing performance measures. The 1x3 and 1x5 blocks of the Inception module help the model learn across the spatial patterns of all dimensional components of the input. A max-pooling layer is added for the removal of noise and elimination of redundant data by downsampling the input data for smaller output. Finally, the max-pooling layer and all the outputs of the convolutional layers are concatenated at filter concatenation. The modified Inception v2 module is shown in Fig 3.

A CBAM block is added between two inception blocks which generate channel and spatial attention highlighting the important parts of the feature map. The concept of CBAM was first introduced by Woo et al. (2018) when the purpose of the authors was to deduce two separate yet sequential attention maps from a given feature map in both channel and spatial dimensions and finally multiply the obtained attention maps with an input map to achieve further feature refinement. Spatial dimensions of the input feature map are compressed in order to achieve effective channel attention. The channel, as well as the spatial attention maps, work complementary to each other. The final output of the Inception block is the input feature map (F) of the CBAM.

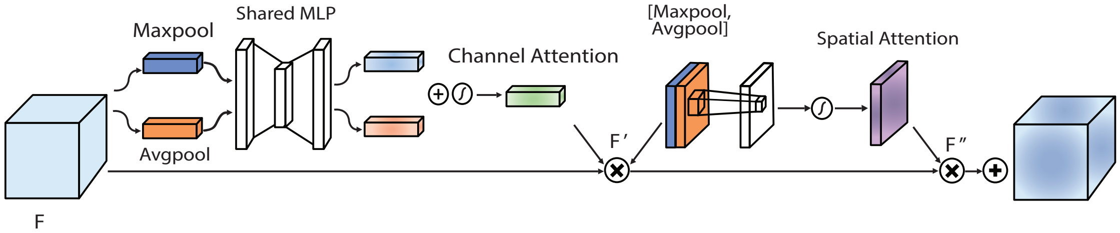


Figure 4: Block diagram of CBAM Woo et al. (2018)

$$M_c(F) = \sigma(MLP(MaxPool(F)) + MLP(AvgPool(F))) \tag{9}$$

$$= \sigma(W_1(W_0(F^c_{max})) + W_1(W_0(F^c_{avg}))) \tag{10}$$

$$F' = M_C(F) * F \tag{11}$$

Here * denotes the element-wise multiplication.

From the block diagram of Fig 4 the working process of CBAM can be observed. For the channel attention map, the spatial information of the input feature map F is aggregated using both average-pooling and max-pooling operations, yielding two distinct intermediate feature maps: $F^s{}_{avg}$ and $F^s{}_{max}$. Both of the maps are forwarded to a shared network consisting of one hidden layer of multi-layer perceptron (MLP) (9) and (10). Here W_0 and W_1 denote the weights of the MLP. After applying the shared network to each feature map, element-wise summation is used to combine the resultant feature vectors (11).

$$M_s(F) = \sigma(f^{7X7}(MaxPool(F')); (AvgPool(F'))) \tag{12}$$

$$= \sigma(f^{7X7}([F^s_{max}; F^s_{avg}])) \tag{13}$$

$$F'' = M_S(F') * F' \tag{14}$$

In order to build an effective feature descriptor for the spatial attention map, average-pooling and max-pooling operations along the channel axis are applied to F′ and concatenated. On the concatenated feature descriptor, a spatial attention map $M_s(F')$ is generated by applying a 7x7 convolutional layer (12) and (13). The resultant feature vector and F′ are combined using element-wise summation to generate the final output F″ (14).

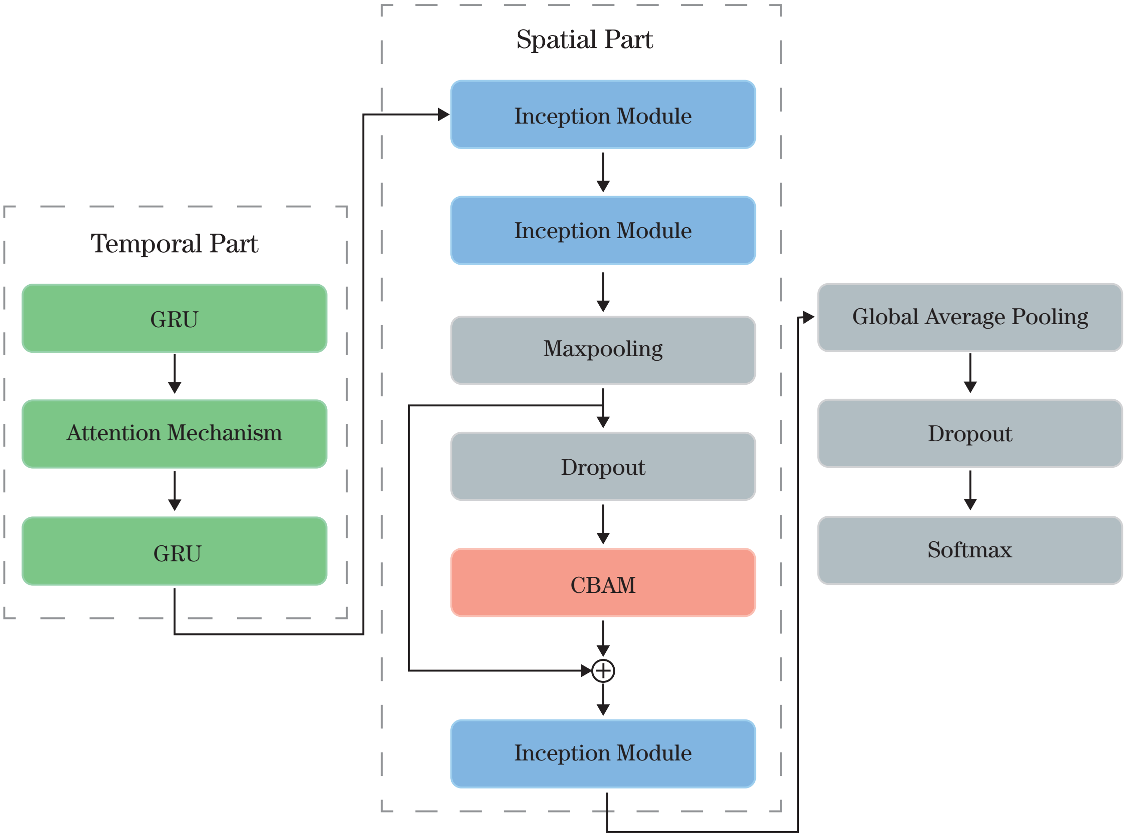


Figure 5: Architecture of the proposed model

The final output of the CBAM block is further passed through another Inception block using a residual connection followed by a GAP layer and a Dropout layer. A residual connection is created, which concatenates the feature maps of Inception and CBAM blocks and prevents performance degradation due to the vanishing gradient problem. The GAP layer is added, replacing the fully connected layers in order to deduct the parameters and the model to converge faster. The Dropout layer reduces the overfitting tendency of the model. Lastly, a Softmax layer is added for classification.

4. Results and Discussion

Although there are a few metrics based on which evaluation can be done, F1-score is chosen for the analysis of the model performance. Since the datasets consist of imbalanced class distribution, other metrics like accuracy do not provide an accurate idea about the recognition rate. On the other hand, F1-score calculates the weighted average of the precision and recall, thus providing a more accurate result.

4.1. Performance Analysis Based on Model Structure

4.1.1. Experiments on Temporal-part:

The proposed model has used an attention mechanism in between two GRU blocks in order to better extract temporal features. The results of an experiment performed on two models with and without Attention Mechanism are shown in Table 2. The model integrating Attention Mechanism exhibits a 1.29% gain in F1-score and a modest reduction in loss, indicating improved performance.

4.1.2. Experiments on Spatial-part:

After a couple of experiments, the proposed model has included a CBAM block for spatial feature extraction. Table 3 depicts the outcomes of the experiments conducted to examine how CBAM improves performance. Although the model containing CBAM has more parameters, its F1-score is 0.85% higher.

Model	F1-score	Loss	No. of Parameters
GRU(144) +GRU(128)	92.85	0.5925	270,486
GRU(144) +AM +GRU(128)	94.14	0.4357	270,758

Table 2: Performance analysis of model architectures with and without Attention Mechanism

Model	F1-score	Loss	No. of Parameters
Inception(192) + Inception(176) + Inception(128)	91.97	1.29	183,702
Inception(192) + Inception(176) + CBAM + Inception(128)	92.82	1.04	206,794

Table 3: Performance analysis of model architectures with and without CBAM

4.1.3. Experiments on GAP layer:

In the proposed model, a GAP layer is added on top of the feature maps instead of fully connected layers. Fully connected layers are connected with the nodes of the upper layers causing weight parameters to occupy more space which is prone to overfitting. On the other hand, GAP layer outputs the average of each feature map which is fed directly into the softmax layer, realizing the goal of decreasing the number of model parameters and reducing computational time thus it enables the model to converge faster. Replacing fully connected layers with GAP layer causes significant change, which can be observed in Table 4.

Model	F1-score%	Loss	No. of Parameters	Computational speed (ms/epoch)
Temporal-part +Spatial-part +FC layer	90.16	0.5542	1,485,254	5,426
Temporal-part +Spatial-part +GAP	96.27	0.133	666,122	4,185

Table 4: Performance analysis of model architectures with fully connected layer and GAP layer

From the results of the experiments, it can be observed that the replacement of GAP layer in the place of Fully connected layers reduces parameters by almost 55.15% and computational speed has also improved by almost 641 ms/epoch. The practicality of using GAP layer is more reasonable as the higher efficacy of the model does not depreciate the performance. Therefore, the model attains an accuracy of 96.27% with the addition of GAP layer.

4.2. Performance Analysis on Datasets

The model is evaluated on five publicly available datasets: UCI-HAR, OPPORTUNITY, PAMAP2, WISDM, and Daphnet to prove its generalization.

4.2.1. UCI-HAR:

Garcia-Gonzalez, Rivero, Fernandez-Blanco & Luaces (2020) proposed the UCI-HAR dataset collected from 30 subjects. The 6 ADL signals were extracted from smartphones' embedded accelerometers and gyroscopes at a sampling frequency rate of 50 Hz. Extracted sensor data were filtered for de-noising purposes and then sampled using 50% overlap and 2.56 sec window size of 128 readings. The dataset is subject-wise divided into train and test datasets consisting of 70% and 30% of the total dataset, respectively.

Activity	Precision	Recall	F1-Score	Support
Walking	0.996644	1.000000	0.998319	297
Walking upstairs	0.965398	0.985866	0.975524	283
Walking downstairs	0.987805	0.964286	0.975904	252
Sitting	0.876161	0.959322	0.915858	295
Standing	0.958763	0.874608	0.914754	319
Laying	1.000000	0.996894	0.998445	322
Weighted Avg.	**0.964056**	**0.962670**	**0.962662**	**1768**

Table 5: Activity-wise precision and recall rate of UCI-HAR dataset

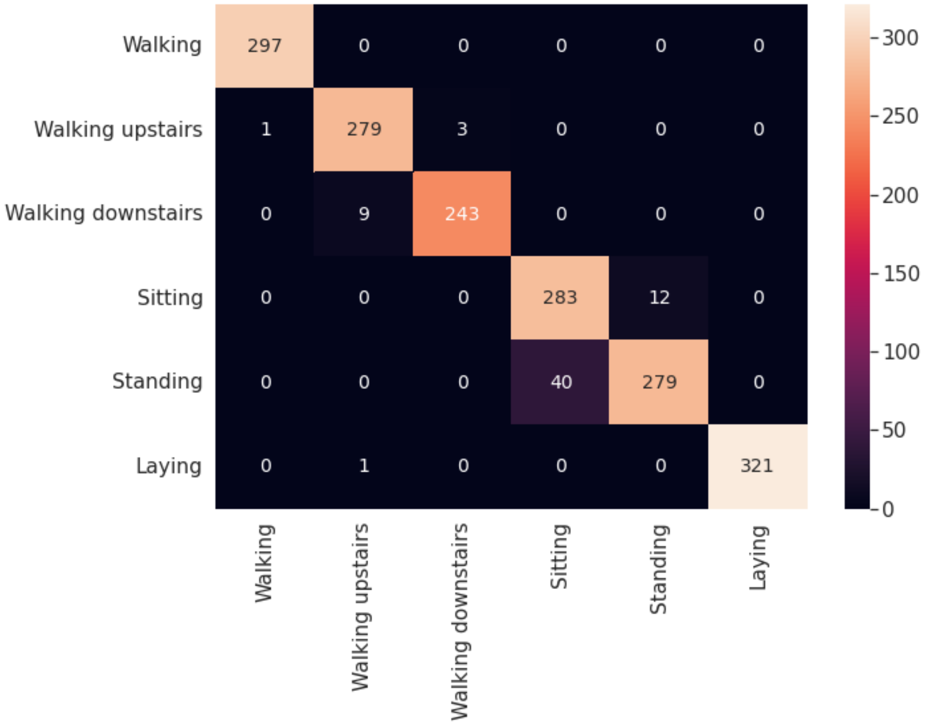


Figure 6: Confusion matrix of UCI-HAR dataset

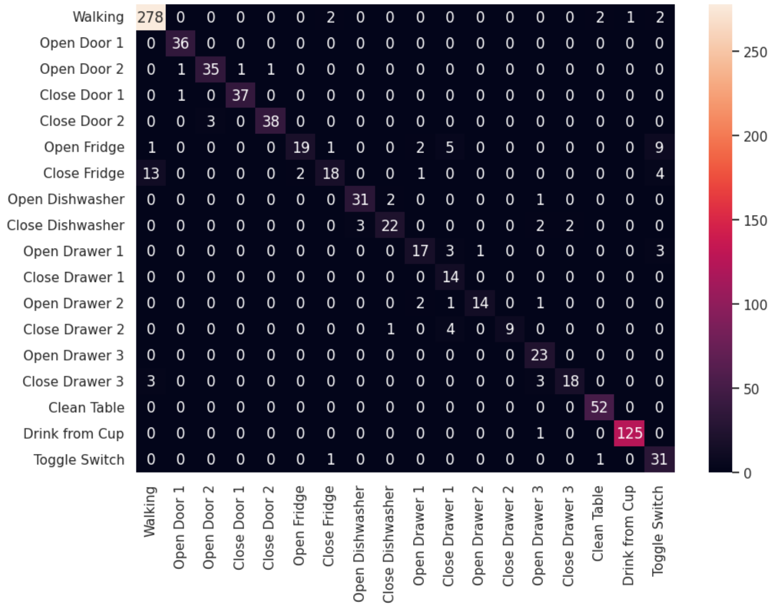
For the UCI-HAR dataset from Table 5, the proposed GRU-INC model achieves 96.27% in terms of weighted average F1-score by acquiring a higher score on almost all the activities. The highest precision and recall rate for the "Laying" activity, the highest recall rate for "Walking" and "Walking upstairs" activities are obtained. "Sitting" and "Standing" activity recognition is observed slightly lower due to the similarity. Fig 6 shows the activity-wise confusion matrix.

4.2.2. OPPORTUNITY:

The OPPORTUNITY dataset was collected from Roggen, Calatroni, Rossi, Holleczek, Förster, Tröster, Lukowicz, Bannach, Pirkl, Ferscha, Doppler, Holzmann, Kurz, Holl, Chavarriaga, Sagha, Bayati, Creatura & Millàn (2010) using 18 complex activities carried out by 4 subjects wearing sensors. Sampled at a rate of 30 Hz, the data used 20% overlap in a fixed-length sliding window of 3 seconds resulting in 90 readings per window. For each subject, five Activities of Daily Living (ADL) and one drill run were collected. The data files were split such as the following were used for

Activity	Precision	Recall	F1-Score	Support
Walking	0.942373	0.975439	0.958621	285
Open Door1	0.947368	1.000000	0.972973	36
Open Door 2	0.921053	0.921053	0.921053	38
Close Door 1	0.973684	0.973684	0.973684	38
Close Door 2	0.974359	0.926829	0.950000	41
Open Fridge	0.904762	0.513514	0.655172	37
Close Fridge	0.818182	0.473684	0.600000	38
Open Dishwasher	0.911765	0.911765	0.911765	34
Close Dishwasher	0.880000	0.758621	0.814815	29
Open Drawer 1	0.772727	0.708333	0.739130	24
Close Drawer 1	0.518519	1.000000	0.682927	14
Open Drawer 2	0.933333	0.777778	0.848485	18
Close Drawer 2	1.000000	0.642857	0.782609	14
Open Drawer 3	0.741935	1.000000	0.851852	23
Close Drawer 3	0.900000	0.750000	0.818182	24
Clean Table	0.945455	1.000000	0.971963	52
Drink from Cup	0.992063	0.992063	0.992063	126
Toggle Switch	0.632653	0.939394	0.756098	33
Weighted Avg.	**0.913748**	**0.903761**	**0.900560**	**904**

Table 6: Activity-wise precision and recall rate of OPPORTUNITY dataset



	Walking	Open Door 1	Open Door 2	Close Door 1	Close Door 2	Open Fridge	Close Fridge	Open Dishwasher	Close Dishwasher	Open Drawer 1	Close Drawer 1	Open Drawer 2	Close Drawer 2	Open Drawer 3	Close Drawer 3	Clean Table	Drink from Cup	Toggle Switch
Walking	278	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	1	2
Open Door 1	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Open Door 2	0	1	35	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Close Door 1	0	1	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Close Door 2	0	0	3	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0
Open Fridge	1	0	0	0	0	19	1	0	0	2	5	0	0	0	0	0	0	9
Close Fridge	13	0	0	0	0	2	18	0	0	1	0	0	0	0	0	0	0	4
Open Dishwasher	0	0	0	0	0	0	0	31	2	0	0	0	0	1	0	0	0	0
Close Dishwasher	0	0	0	0	0	0	0	3	22	0	0	0	0	2	2	0	0	0
Open Drawer 1	0	0	0	0	0	0	0	0	0	17	3	1	0	0	0	0	0	3
Close Drawer 1	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0
Open Drawer 2	0	0	0	0	0	0	0	0	0	2	1	14	0	1	0	0	0	0
Close Drawer 2	0	0	0	0	0	0	0	0	1	0	4	0	9	0	0	0	0	0
Open Drawer 3	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0
Close Drawer 3	3	0	0	0	0	0	0	0	0	0	0	0	0	3	18	0	0	0
Clean Table	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0
Drink from Cup	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	125	0
Toggle Switch	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	31

Figure 7: Confusion matrix of OPPORTUNITY dataset

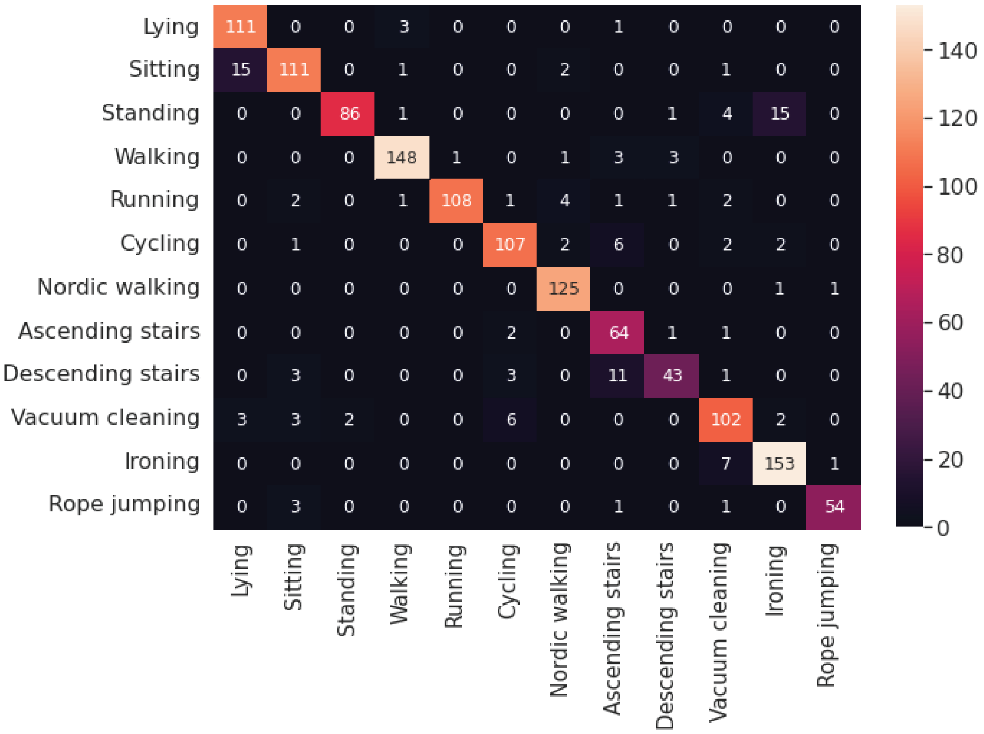
testing: ’Run 5 of Subject 1’, ’Drill of Subject 2’, ’Run 1 and 4 of Subject 3’ and ’Run 4 of Subject 4’. The others were selected for training and validation.

Evaluating subject-wise and considering the IMU sensors, the proposed model achieves 90.05% F1-score. The mixed performance can be explained as such, the model achieves 100% recall on four kinds of activities which are "Open Door 1", "Close Drawer 1", "Open Drawer 3," and "Clean Table." However, comparatively low recall but good precision on the activities like “Open Fridge” and “Close Fridge” can be observed from Table 6. Fig 7 depicts the confusion matrix.

4.2.3. PAMAP2:

Activity	Precision	Recall	F1-Score	Support
Lying	0.860465	0.965217	0.909836	115
Sitting	0.902439	0.853846	0.877470	130
Standing	0.977273	0.803738	0.882051	107
Walking	0.961039	0.948718	0.954839	156
Running	0.990826	0.900000	0.943231	120
Cycling	0.899160	0.891667	0.895397	120
Nordic walking	0.932836	0.984252	0.957854	127
Ascending stairs	0.735632	0.941176	0.825806	68
Descending stairs	0.877551	0.704918	0.781818	61
Vacuum cleaning	0.842975	0.864407	0.853556	118
Ironing	0.884393	0.950311	0.916168	161
Rope jumping	0.964286	0.915254	0.939130	59
Weighted Avg.	**0.907849**	**0.903130**	**0.903078**	**1342**

Table 7: Activity-wise precision and recall rate of PAMAP2 dataset



	Lying	Sitting	Standing	Walking	Running	Cycling	Nordic walking	Ascending stairs	Descending stairs	Vacuum cleaning	Ironing	Rope jumping
Lying	111	0	0	3	0	0	0	1	0	0	0	0
Sitting	15	111	0	1	0	0	2	0	0	1	0	0
Standing	0	0	86	1	0	0	0	0	1	4	15	0
Walking	0	0	0	148	1	0	1	3	3	0	0	0
Running	0	2	0	1	108	1	4	1	1	2	0	0
Cycling	0	1	0	0	0	107	2	6	0	2	2	0
Nordic walking	0	0	0	0	0	0	125	0	0	0	1	1
Ascending stairs	0	0	0	0	0	2	0	64	1	1	0	0
Descending stairs	0	3	0	0	0	3	0	11	43	1	0	0
Vacuum cleaning	3	3	2	0	0	6	0	0	0	102	2	0
Ironing	0	0	0	0	0	0	0	0	0	7	153	1
Rope jumping	0	3	0	0	0	0	0	1	0	1	0	54

Figure 8: Confusion matrix of PAMAP2 dataset

Reiss & Stricker (2012) proposed the Physical Activity Monitoring (PAMAP2) dataset consisting of 18 physical activities of daily life performed by 9 subjects. Subject 5 was selected for the test dataset and the remaining data for the train and validation datasets. The subjects wore sensors on their hands, chest and ankles. The sensors included an accelerometer, gyroscope, magnetometer, and heart rate monitor. The data were sampled at a rate of 100 Hz using 50% overlap with a fixed-length sliding window resulting in 256 readings per window. From Table 7, for the PAMAP2 dataset, the proposed model achieves a 90.30% F1-score whereas the "Nordic walking" achieves the highest F1-score of 95.78%. In the case of precision, the best-performing activity is "Running", and the least precision is observed in the "Ascending Stairs" activity and the least recall is observed in the "Descending stairs" activity. Fig 8 displays the confusion matrix which clearly depicts the recognition accuracy.

4.2.4. WISDM:

The WISDM dataset proposed by Kwapisz, Weiss & Moore (2010) comprised data collected from 36 test subjects performing 6 ADL with accelerometer and gyroscope embedded in smartphones and smartwatches at a sampling rate of 20 Hz.

Activity	Precision	Recall	F1-Score	Support
Walking	0.989510	1.000000	0.994728	849
Jogging	1.000000	1.000000	1.000000	684
Walking Upstairs	0.965241	0.970430	0.967828	372
Walking Downstairs	0.989761	0.953947	0.971524	304
Sitting	1.000000	1.000000	1.000000	363
Standing	1.000000	1.000000	1.000000	293
Weighted Avg.	**0.991292**	**0.991274**	**0.991239**	**2865**

Table 8: Activity-wise precision and recall rate of WISDM dataset

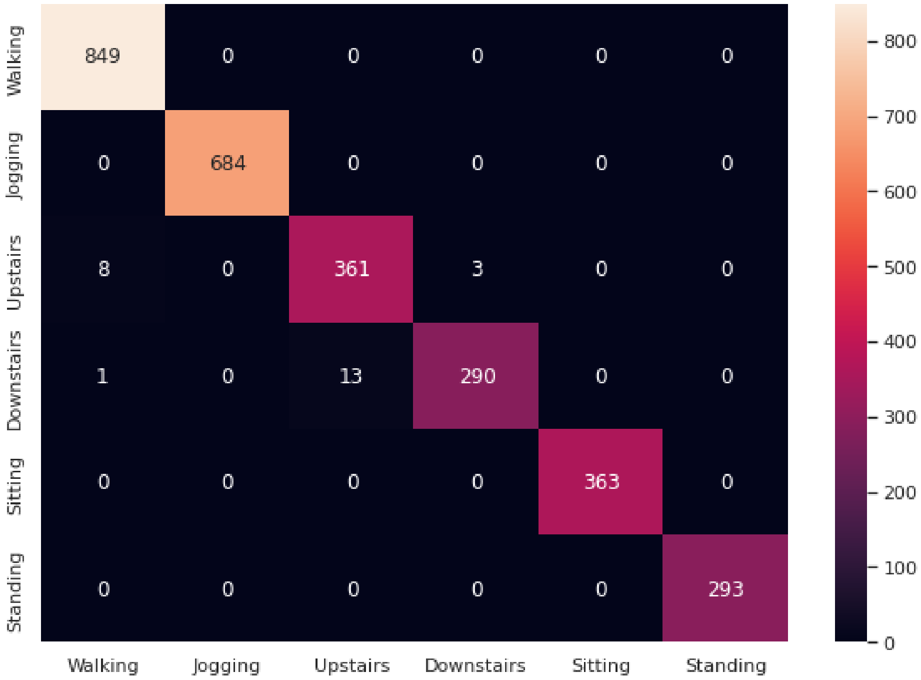


Figure 9: Confusion matrix of WISDM dataset

Only choosing the simple activities proves that the model can successfully detect complex as well as simple activities, justifying the generalization capability of the model. In this dataset, the proposed GRU-INC achieves significantly high performance on all parameters, which can be observed in Table 8. It achieves about 99.12% precision, recall, and F1-Score on all activities. A 100% F1-Score is achieved in almost all activities with the exceptions of the following activities: "Walking", "Walking Upstairs," and "Walking Downstairs". Fig 9 represents the confusion matrix of the GRU-INC model on this dataset.

4.2.5. Daphnet:

Bächlin, Plotnik, Roggen, Giladi, Hausdorff & Tröster (2009) proposed the Daphnet dataset, which contains only two classes labeled as "Freeze" or "No Freeze" classes sampled from 10 users. The data was sampled using 50% overlap and a 3-sec window size of 192 readings.

Activity	Precision	Recall	F1-Score	Support
No freeze	0.971311	0.948000	0.959514	250
Freeze	0.949219	0.972000	0.960474	250
Weighted Avg.	**0.960265**	**0.960000**	**0.959994**	**500**

Table 9: Activity-wise precision and recall rate of Daphnet dataset

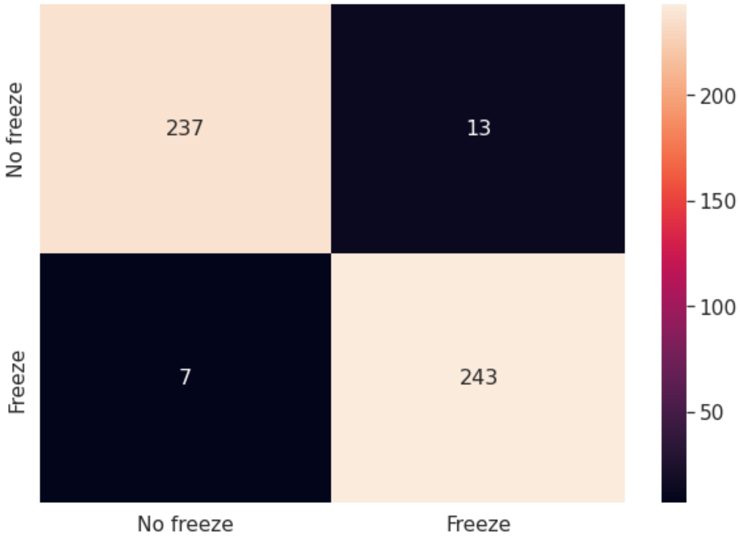


Figure 10: Confusion matrix of Daphnet dataset

The model achieved an F1-score of 95.99% with a 97.2% precision rate for the "Freeze" class. The corresponding confusion matrix can be observed in Fig 10 where the recognition rate can be observed more precisely.

4.3. Comparative Analysis

The proposed model has been compared with baseline models and some existing works based on F1-score, loss, and the number of parameters to justify its effectiveness. The analysis shows that the GRU-INC model has achieved higher accuracy and minimal loss as well as reduced computational complexity due to its comparatively wider structure rather than deep.

4.3.1. Comparative Analysis with Baseline Models

For proper evaluation, this study has chosen the following baseline models: CNN, Vanilla Long Short-Term Memory (vLSTM), BiLSTM, Bi-directional Gated Recurrent Unit (BiGRU), and Deep Convolutional Long Short-Term Memory (DeepConvLSTM). The study experimented with each of these models in the chosen five datasets and compared its results with the proposed model. The comparison in terms of F1-score and loss are shown in Table 10 and Table 11 respectively.

Model	UCI-HAR	OPPORTUNITY	PAMAP2	WISDM	Daphnet
CNN (Yang, Nguyen, San, Li & Krishnaswamy, 2015)	92.39	81.74	85.31	95.70	91.37
vLSTM (Mutegeki & Han, 2020)	91.15	82.52	85.19	95.61	91.79
BiLSTM (Wan, Qi, Xu, Tong & Gu, 2020)	90.90	83.03	83.03	96.96	93.59
BiGRU (Wan et al., 2020)	90.15	82.55	86.65	97.86	92.96
DeepConvLSTM (Ordóñez & Roggen, 2016)	92.54	81.83	87.37	97.13	93.78
GRU-INC	**96.27**	**90.05**	**90.30**	**99.12**	**95.99**

Table 10: Comparative analysis among the proposed model with baseline models in terms of F1-score

Model	UCI-HAR	OPPORTUNITY	PAMAP2	WISDM	Daphnet
CNN (Yang et al., 2015)	0.424	0.741	0.694	0.111	0.230
vLSTM (Mutegeki & Han, 2020)	0.268	0.811	0.635	0.118	0.241
BiLSTM (Wan et al., 2020)	0.293	0.770	0.519	0.112	0.207
BiGRU (Wan et al., 2020)	0.314	0.662	0.498	0.098	0.396
DeepConvLSTM (Ordóñez & Roggen, 2016)	0.469	0.694	0.617	0.096	0.212
GRU-INC	**0.133**	**0.389**	**0.428**	**0.044**	**0.188**

Table 11: Comparative analysis among the proposed model with baseline models in terms of loss

DeepConvLSTM has the highest accuracy among the baseline models on UCI-HAR, PAMAP2, and Daphnet datasets. The proposed GRU-INC model performs significantly better even than DeepConvLSTM by achieving 3.73%, 2.93%, and 2.21% higher accuracy of F1-score as well as lower loss on the above-mentioned three datasets. Although the OPPORTUNITY dataset consists of complex activities, the proposed model can better detect the activities compared with other baseline models due to the usage of the Attention Mechanism separately in both the temporal and spatial parts of the model. It can be observed from Table 10 that BiLSTM has the highest accuracy of the other baseline models for the OPPORTUNITY dataset. The GRU-INC model has 7.02% higher accuracy than that of the BiLSTM and, at the same time, 0.381 lesser loss. The GRU-INC model achieves an F1-Score of 99.12% on the WISDM dataset, which is higher than BiGRU, the best-performed baseline model on this dataset. The loss on the GRU-INC model is comparatively lower than all the other models.

4.3.2. Comparative Analysis with Existing Models

A further experiment was conducted by keeping similar data distribution to compare the proposed GRU-INC model with some existing models based on F1-score and the number of parameters where it outperformed all the other models. The comparison based on UCI-HAR, WISDM, and Daphnet datasets is shown in Table 12.

Dataset	Model	F1-score/ Acc.(%)	No. of Parameters
UCI-HAR	iSPLInception (Ronald et al., 2021)	95	1,338,651
	LSTM-CNN (Xia et al., 2020)	95.78	-
	InnoHAR (Xu et al., 2019)	94.5	-
	GRU-INC (proposed)	**96.27**	**666,122**
WISDM	DanHAR (Gao et al., 2021)	98.85	2.32M
	LSTM-CNN (Xia et al., 2020)	95.85	-
	GRU-INC (proposed)	**99.12**	**661,486**
Daphnet	iSPLInception (Ronald et al., 2021)	94	1,326,726
	GRU-INC (proposed)	**95.99**	**665,670**

Table 12: Comparative analysis among the proposed model with existing models based on UCI-HAR, WISDM, and Daphnet datasets

	Dataset	Model	F1-score/ Acc.(%)	No. of Parameters
Subject-wise	OPPORTUNITY	iSPLInception (Ronald et al., 2021)	88	1,354,789
		DanHAR (Gao et al., 2021)	82.75	1.57M
		GRU-INC (proposed)	**90.05**	**723,728**
Non Subject-wise	OPPORTUNITY	LSTM-CNN (Xia et al., 2020)	92.63	-
		InnoHAR (Xu et al., 2019)	94.5	-
		GRU-INC (proposed)	**96.15**	**723,728**
Subject-wise	PAMAP2	iSPLInception (Ronald et al., 2021)	89.09	1,338,651
		DanHAR (Gao et al., 2021)	93.15	3.51M
		AttnSense (Ma et al., 2019)	89.3	-
		GRU-INC (proposed)	**90.31**	**642,800**
Non Subject-wise	PAMAP2	InnoHAR (Xu et al., 2019)	93.5	-
		GRU-INC (proposed)	**95.61**	**723,728**

Table 13: Comparative analysis among the proposed model with existing models based on OPPORTUNITY and PAMAP2 datasets

It can be observed that for the UCI-HAR dataset, the proposed model has an accuracy of 96.27%, which is 1.27% higher than the iSPLInception (Ronald et al., 2021) model's accuracy, although they used more than twice the number of the parameters compared to that of the proposed model. Again for the WISDM dataset, the proposed model has a close accuracy of 0.28% higher than the DanHAR (Gao et al., 2021) model, which is highly expensive as it uses almost four times greater parameters than the proposed one. For Daphnet dataset, GRU-INC model achieves 1.99% more accuracy than that of the iSPLInception (Ronald et al., 2021) model.

For the OPPORTUNITY and PAMAP2 datasets, the experiment has been performed for both subject-wise and non-subject-wise data distribution. The comparison based on these two datasets is shown in Table 13. It can be noted that the proposed model exceeds the other existing models as it can detect complex activities more accurately because of the incorporation of attention mechanisms individually in temporal and spatial parts. For subject-wise data distribution in the OPPORTUNITY dataset, the proposed model has achieved 2.05% and 7.3% higher accuracy than iSPLInception (Ronald et al., 2021) and DanHAR, (Gao et al., 2021) respectively. For the PAMAP2 dataset, the GRU-INC model shows a higher accuracy of 90.31% compared with almost all the other existing works in Table 13 except DanHAR, (Gao et al., 2021) which secures 2.84% more accuracy. As DanHAR (Gao et al., 2021) used almost six times more parameters compared to that of the proposed model, and thus it achieved higher accuracy. However, the GRU-INC model performs efficiently and significantly due to the minimal computational complexity. Again, for non-subject-wise data distribution, the proposed model surpasses the other existing models for both datasets.

5. Conclusion

GRU-INC, is a novel approach for HAR which uses GRU and Inception module along with Attention mechanisms to identify complex human activities. The inception module is used for better performance and to reduce the number of parameters. To further refine the features extracted from both temporal and spatial data, a CBAM block is added and incorporated with the inception module. The model is made more efficient by replacing the fully connected layer with a GAP layer to reduce model parameters, thus enabling faster convergence of the model. In this paper, the proposed model proves to be generalized by attaining a high F1-score and lower loss being evaluated on five publicly available datasets such as UCI-HAR, OPPORTUNITY, PAMAP2, WISDM, and Daphnet. The adoption of attention mechanisms separately in both the spatial and temporal parts has made GRU-INC capable of detecting the challenging human activities, which mostly comprise the OPPORTUNITY and PAMAP2 datasets. F1-score of 90.05% and 90.31% have been respectively achieved in these datasets. The proposed model is compared with some baseline models as well as some existing approaches where it demonstrates a higher F1-score than each of these models. Although the proposed model showed an overall good performance rate in determining the complex activities in the OPPORTUNITY dataset, it acquired a low F1-score in distinguishing a few activities which are similar in type. Additionally, a better recognition rate for complex activities can be achieved if the sensor-wise concentration for different features is provided to the activities.

Acknowledgement: The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding and supporting this work through Research Partnership Program no RP-21-09-09 .

References

Abbaspour, S., Fotouhi, F., Sedaghatbaf, A., Fotouhi, H., Vahabi, M., & Linden, M. (2020). A comparative analysis of hybrid deep learning models for human activity recognition. *Sensors*, *20*, 5707.

Abdel-Basset, M., Hawash, H., Chakrabortty, R. K., Ryan, M., Elhoseny, M., & Song, H. (2020). St-deephar: Deep learning model for human activity recognition in ioht applications. *IEEE Internet of Things Journal*, *8*, 4969–4979.

Alema Khatun, M., & Abu Yousuf, M. (2020). Human activity recognition using smartphone sensor based on selective classifiers. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)* (pp. 1–6). doi:10.1109/STI50764.2020.9350486.

Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B., & Kechadi, M. T. (2019). Human activity recognition with convolutional neural netowrks. *CoRR*, *abs/1906.01935*.

Buffelli, D., & Vandin, F. (2021). Attention-based deep learning framework for human activity recognition with user adaptation. *IEEE Sensors Journal*, .

Bächlin, M., Plotnik, M., Roggen, D., Giladi, N., Hausdorff, J., & Tröster, G. (2009). A wearable system to assist walking of parkinson´s disease patients. *Methods of information in medicine*, *49*, 88–95. doi:10.3414/ME09-02-0003.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, .

Ehatisham-Ul-Haq, M., Javed, A., Azam, M. A., Malik, H. M. A., Irtaza, A., Lee, I. H., & Mahmood, M. T. (2019). Robust human activity recognition using multimodal feature-level fusion. *IEEE Access*, *7*, 60736–60751. doi:10.1109/ACCESS.2019.2913393.

Gao, W., Zhang, L., Teng, Q., He, J., & Wu, H. (2021). Danhar: Dual attention network for multimodal human activity recognition using wearable sensors. *Applied Soft Computing*, *111*, 107728.

Garcia-Gonzalez, D., Rivero, D., Fernandez-Blanco, E., & Luaces, M. R. (2020). A public domain dataset for real-life human activity recognition using smartphone sensors. *Sensors*, *20*. doi:10.3390/s20082200.

Ghazal, S., Khan, U. S., Mubasher Saleem, M., Rashid, N., & Iqbal, J. (2019). Human activity recognition using 2d skeleton data and supervised machine learning. *IET Image Processing*, *13*, 2572–2578. doi:https://doi.org/10.1049/iet-ipr.2019.0030.

Haque, M. N., Tonmoy, M. T. H., Mahmud, S., Ali, A. A., Khan, M. A. H., & Shoyaib, M. (2019). Gru-based attention mechanism for human activity recognition. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (pp. 1–6). IEEE.

Hassan, M. M., Uddin, M. Z., Mohamed, A., & Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, *81*, 307–313. doi:https://doi.org/10.1016/j.future.2017.11.029.

Hernández, F., Suárez, L. F., Villamizar, J., & Altuve, M. (2019). Human activity recognition on smartphones using a bidirectional lstm network. In *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)* (pp. 1–5). doi:10.1109/STSIVA.2019.8730249.

Kwapisz, J., Weiss, G., & Moore, S. (2010). Activity recognition using cell phone accelerometers. *SIGKDD Explorations*, *12*, 74–82. doi:10.1145/1964897.1964918.

Ma, H., Li, W., Zhang, X., Gao, S., & Lu, S. (2019). Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI* (pp. 3109–3115).

Mekruksavanich, S., & Jitpattanakul, A. (2021). Lstm networks using smartphone data for sensor-based human activity recognition in smart homes. *Sensors*, *21*. doi:10.3390/s21051636.

Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., & Dürichen, R. (2017). Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* ISWC '17 (p. 158–165). Association for Computing Machinery. doi:10.1145/3123021.3123046.

Mutegeki, R., & Han, D. S. (2020). A cnn-lstm approach to human activity recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 362–366). IEEE.

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

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, *16*, 115.

Qi, W., Su, H., Yang, C., Ferrigno, G., De Momi, E., & Aliverti, A. (2019). A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone. *Sensors*, *19*, 3731.

Reiss, A., & Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers* (pp. 108–109). doi:10.1109/ISWC.2012.13.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., Doppler, J., Holzmann, C., Kurz, M., Holl, G., Chavarriaga, R., Sagha, H., Bayati, H., Creatura, M., & Millàn, J. d. R. (2010). Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh International Conference on Networked Sensing Systems (INSS)* (pp. 233–240). doi:10.1109/INSS.2010.5573462.

Ronald, M., Poulose, A., & Han, D. S. (2021). isplinception: an inception-resnet deep learning architecture for human activity recognition. *IEEE Access*, *9*, 68985–69001.

Schrader, L., Toro, A., Konietzny, S., Rüping, S., Schäpers, B., Steinböck, M., Krewer, C., Mueller, F., Guettler, J., & Bock, T. (2020). Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing*, *13*. doi:10.1007/s12062-020-09260-z.

Shojaedini, S. V., & Beirami, M. J. (2020). Mobile sensor based human activity recognition: distinguishing of challenging activities by applying long short-term memory deep learning modified by residual network concept. *Biomedical Engineering Letters*, *10*, 419–430.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Taylor, W., Shah, S. A., Dashtipour, K., Zahid, A., Abbasi, Q. H., & Imran, M. A. (2020). An intelligent non-invasive real-time human activity recognition system for next-generation healthcare. *Sensors*, *20*. doi:10.3390/s20092653.

Um, T. T., Pfister, F. M. J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., & Kulić, D. (2017). Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* ICMI '17 (p. 216–220). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3136755.3136817.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, *25*, 743–755.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).

Xia, K., Huang, J., & Wang, H. (2020). Lstm-cnn architecture for human activity recognition. *IEEE Access*, *8*, 56855–56866.

Xu, C., Chai, D., He, J., Zhang, X., & Duan, S. (2019). Innohar: A deep neural network for complex human activity recognition. *Ieee Access*, *7*, 9893–9902.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial intelligence*.

Yao, S., Hu, S., Zhao, Y., Zhang, A., & Abdelzaher, T. (2017). Deepsense: A unified deep learning framework for time-series mobile sensing data processing. arXiv:1611.01942.

Zhang, B., Xu, H., Xiong, H., Sun, X., Shi, L., Fan, S., & Li, J. (2020). A spatiotemporal multi-feature extraction framework with space and channel based squeeze-and-excitation blocks for human activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, (pp. 1–13).

Zhang, H., Xiao, Z., Wang, J., Li, F., & Szczerbicki, E. (2019). A novel iot-perceptive human activity recognition (har) approach using multihead convolutional attention. *IEEE Internet of Things Journal*, *7*, 1072–1080.

Zhang, J., Qiao, S., Lin, Z., & Zhou, Y. (2021). Human activity recognition based on residual network. In *IOP Conference Series: Earth and Environmental Science* (p. 012041). IOP Publishing volume 693.

Zhou, X., Liang, W., Wang, K. I.-K., Wang, H., Yang, L. T., & Jin, Q. (2020). Deep-learning-enhanced human activity recognition for internet of healthcare things. *IEEE Internet of Things Journal*, *7*, 6429–6438. doi:10.1109/JIOT.2020.2985082.

View publication stats