International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Oct-2022)

https://dx.doi.org/10.12785/ijcds/120188

Comprehensive Analysis of Deep Learning-based Human Activity Recognition approaches based on Accuracy

Aniruddh G. Fataniya[1] and Dr. Hardik P. Modi[2]

[1]*Department of Computer Engineering, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India*

[2]*Department of Electronics and Communication Engineering, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India*

Received 22 Jan. 2021, Revised 15 Jul. 2022, Accepted 23 Jul. 2022, Published 31 Oct. 2022

Abstract: Human Activity Recognition (HAR) is a vital area of Computer Vision. HAR focuses on various activities carried out by humans. Information relative to the human activities is collected by smart sensors and wearable devices. HAR is classified into two categories, e.g. (a) Vision-based, i.e. human activities are captured in form of image and video and (b) Sensor-based, i.e. human activity input can be taken from wearable devices and object tagging techniques. Human activity recognition is an extensive thrust area for Content-based video analysis, Human-machine interaction, animation, healthcare fields. The paper presents a comprehensive analysis of various deep learning-based approaches adopted to implement human activity recognition based on accuracy. It is observed that for the vision-based category the performance of the Depth Camera-based Recurrent Neural Network model is 99.55% accuracy with 12 activities for MSRC-12 datasets and for the sensor-based category, the performance of HAR by Wearable sensors using Deep Neural Network model is 99.93% accuracy with 03 activities for SHO datasets. It is also observed that for Opportunity dataset, InnoHAR: A DNN for complex HAR model gives good performance with 94.6% accuracy along with 18 activities, for PAMAP2 dataset, Multi-input CNN-GRU model gives good performance with 95.27% accuracy along with 12 activities, for WISDM dataset, ConvAE-LSTM model gives good performance with 98.67% accuracy along with 6 activities, and for UCI-HAR dataset, ConvAE-LSTM model gives good performance with 98.14% accuracy along with 6 activities.

Keywords: Human Activity Recognition (HAR), Deep Learning, Vision-based Human Activity Recognition, Sensor-based Human Activity Recognition

1. Introduction

Human activity recognition (HAR) is a contentious research area in computer vision. It is extremely important in human-to-human interaction and interpersonal relationships. HAR can be defined as the ability to recognize human activity based on information received from various sensors [1]. Cameras, wearable sensors, sensors attached to everyday objects, or sensors deployed in the environment are all examples of sensors.

Different approaches have been used to capture various activities. As illustrated in figure.1, HAR can be divided in two kinds of approaches: vision-based and sensor-based. A camera is used in a vision-based approach to capture information about human activities. Different activities are frequently recognized by using computer vision techniques on this captured data. Although computer vision-based techniques are simple to use and can produce good results, they have a number of drawbacks. The primary concern is privacy. Another issue with this approach is its reliance on light. When there is no light, traditional cameras cannot detect it.

The other approach is sensor-based, in which human behavior is identified using sensors. This approach is further classified into three deployment types: i) wearable, ii) object-tagged, and iii) dense sensing. In the wearable approach, the sensor is carried by the user and detects activity. However, carrying the sensors is not always feasible. Sensors are attached to everyday objects in the Object-tagged approach. This is also not possible because the user is constrained by the tagged-objects. The sensors are deployed in the surroundings in the Dense Sensing approach, and the activities are captured by the sensors when a user performs some activities. This approach appears to be more feasible than the others because the user does not need to carry the sensors or object-tagged devices with them. However, there are still some challenges, such as noise in the environment, which affects the interference in data capture.

E-mail address: aniruddhfataniya.ce@charusat.ac.in, hardikmodi.ec@charusat.ac.in

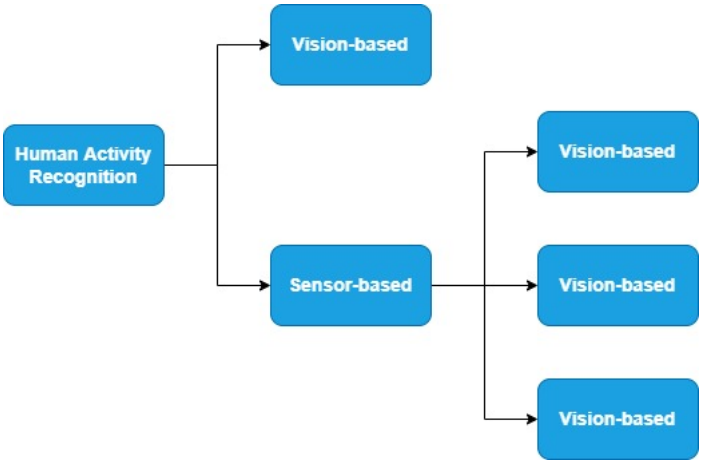

Figure 1. Classification of HAR

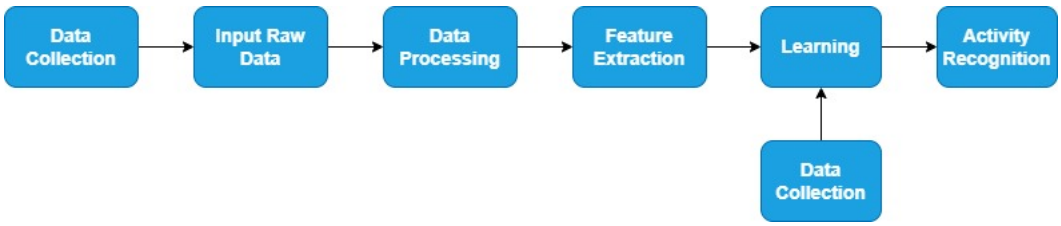

Figure 2. General Architecture of HAR

A general architecture for HAR is depicted in Figure 2. Firstly, data is collected from the sensors and pre-processed based on the requirement of the model used. Next step is feature extraction which extracts important features from the data. In the next step, the model will learn and be trained from the dataset. Finally, the activity is recognized.

In this paper, we have reviewed some research papers in the area of HAR. This review will help the researchers to select the appropriate model in performing Human activity recognition on a particular dataset.

2. Literature Review

CNN-LSTM [2]: Ronald et al. presented CNN-LSTM recognizing and classifying different activities carried out by humans. A mixed approach of CNN and LSTM was used to identify activities.

Functionality of Model: As shown in Figure.3, a 1D convolution layer with the ReLu activation function was used. They then added a flatten layer and a 1D maxpooling layer to format the feature data so that it could be consumed by the LSTM layer in the following step. The data on which the convolution layer operates is very different from the data on which the LSTM layer operates. Keras was used to dealing with this challenge. The LSTM layer was also activated using the ReLu function. The LSTM layer's output was fed into the Softmax-activated fully connected output layer. This layer categorizes the input into the activity's class.

Experiment setup and Used tool:The proposed approach is implemented using Keras with Tensorflow as a back end and validated on iSPL and UCI-HAR datasets.
Comment: The suggested model gives 99.0% accuracy for 3 activities (walking, sitting and standing) for iSPL dataset as compared to other models and datasets.
Limitation: The model could be evaluated on different hyper parameters and against other publicly available datasets.

Acceleration-based HAR using CNN [3]: Yuqing et al. presented a CNN approach to Acceleration-based HAR for human activity recognition. A CNN model with a modified kernel was built to adapt the properties of triaxial acceleration signals.
Functionality of Model: A CNN architecture was modified by changing the angle of the convolution kernel, and a method was used to determine the best number of epochs. The width of the convolution kernel was set to 2 in order to extract information between different axes and improve rotation flexibility. With a kernel width of 2, CNN has a low error rate. A set of samples was chosen for validation in order to select the best epoch, and every 10 epochs, training error rates were evaluated and validated. When the validation error rate stopped decreasing and started increasing, training and evaluating the network on the test set was halted. CNN's architecture consists of three convolution layers and three pooling layers. It is entirely based on the raw acceleration signal, with no additional processing.

Figure 3. Block diagram of the CNN-LSTM architecture for HAR [2]

Experiment setup and Used tool:The proposed approach is validated on Self-made dataset. However, The tool used for the implementation is not mentioned this proposed article.
Comment: The suggested model gives 93.8% accuracy for 8 activities for self-made dataset as compared to other models.
Limitation: The model could be evaluated and validated against some publicly available datasets.

RNN-based HAR [4]: Park et al. represented the approach of RNN-based HAR for recognizing human activities. LSTM was used for this architecture.
Functionality of Model: Human activities were visualized as time-sequenced variation in multiple joint angles. RNN was built using 50 LSTMs and 90 hidden units. Here, the length of the activity video frames is reflected by the number of LSTMs. The model was trained using the Extended Backpropagation algorithm on the training feature data. Figure.4 depicts the RNN-based HAR model, in which a human silhouette is extracted from depth camera data, a 3D human pose is recognized, and time sequential variation in joint angles are given in the LSTM model, which identifies human activity.
Experiment setup and Used tool: The proposed approach is validated on MSRC-12 dataset. However, The tool used for the implementation is not mentioned this proposed article.
Comment: The suggested model gives 99.55% accuracy for 12 activities for MSRC-12 dataset as compared to other models.
Limitation: Datasets related to depth cameras are limited. Processing of Depth Camera images is somewhat complex.

ST-GCN [5]: Xin et al. demonstrated ST-GCN activity recognition using 3D motion data.
Functionality of Model: As shown in figure.5, batch normalization was utilized to normalize the input skeleton data. This model consisted of nine layers consisting units of ST-GCN. All layers have different number of output channels. The temporal kernel size for this layer was set to 9. The residual mechanism and a dropout layer were applied to each unit of ST-GCN to avoid overfitting. Softmax Classifier was used for classification in the end.
Experiment setup and Used tool:The proposed approach is implemented using python and PyTorch framework and trained on two GTX 1080 Ti GPUs and validated on Carecom nurse care activity dataset.
Comment: The suggested model gives 57.0% accuracy for 6 activities for Care Com's nurse care activity.
Limitation: Due to a lack of data, the model's results are unstable.

AttnSense for multimodal human activity recognition [6]: Haojie et al. represented AttSense for multimodal human activity recognition for recognizing human activities. It is the combination of the attention mechanism with CNN and GRU.
Functionality of Model: Data preprocessing includes data augment, fast Fourier transform, and data segmentation. AttSense is made up of three layers: a convolution subnet, an attention-based GRU subnet, and an output layer. The convolution subnet is composed of stacked convolution layers and pooling layers. In addition, at each layer, a batch normalization layer is used to reduce internal covariate shift. A self-attention network was introduced in the attention-fusion subnet, in which sensor feature vectors were fed as input and an attention weight for each modality was output. The importance of various sensors in the HAR job is represented by these attention weights. The feature vectors are combined to form a uniform feature representation vector. The attentionfusion subnet's output was routed to a stacked GRU structure. Gate units transformed the input into hidden layer output. The self-attention mechanism was employed once more to compute the weighted average sum of all hidden states. The output of the attentionbased GRU subnet was fed into the output layer, which calculates the probability of each activity using a fullyconnected Softmax function.
Experiment setup and Used tool:The proposed approach is implemented using Tensorflow and trained on GTX 1070ti GPU and validated on Heterogeneous, Skoda and PAMAP2 datasets.
Comment: The suggested model gives 96.5% accuracy for 6 activities (standing, sitting, biking, walking, , climbup, climb-down) for Heterogeneous dataset as compared to other models and datasets.
Limitation: Multiple sensors are required for different body parts which makes the system complex for design and cost ineffective. An Individual CNN is required for an individual sensor which tends to design complex algorithms.

Binarized-BLSTM-RNN based HAR [7]: Marcus et al. represented Binarized-BLSTM-RNN for recognizing human activities. The model is based on Bidirectional LSTM. The weight parameters, as well as the output signals of the input and in-between hidden layers, are binary-valued, requiring only basic bit logic for training and testing.
Functionality of Model: The binary weights were used to train BLSTM-RNN and LSTM. However, binarized weights were only used during the forward and backward passes, and high precision weights were used in both cases to update the parameters. However, because the changes in gradient descent parameters were so minor, the training objective was not improved. As a result, RNN was trained

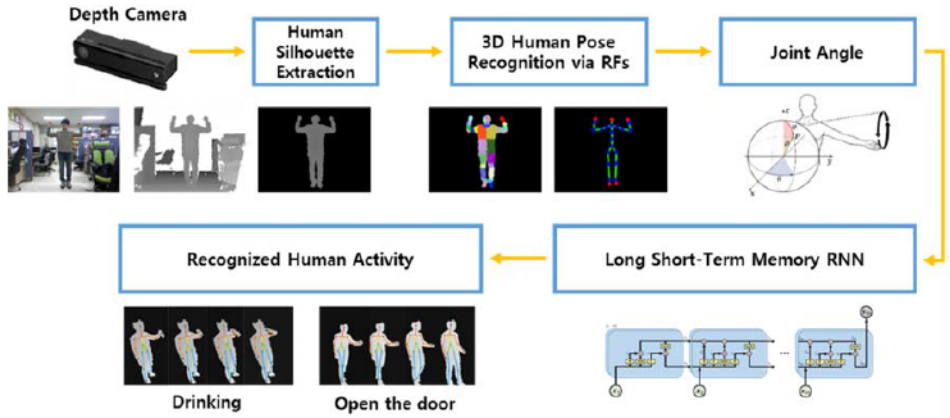

Figure 4. RNN-base HAR System [4]

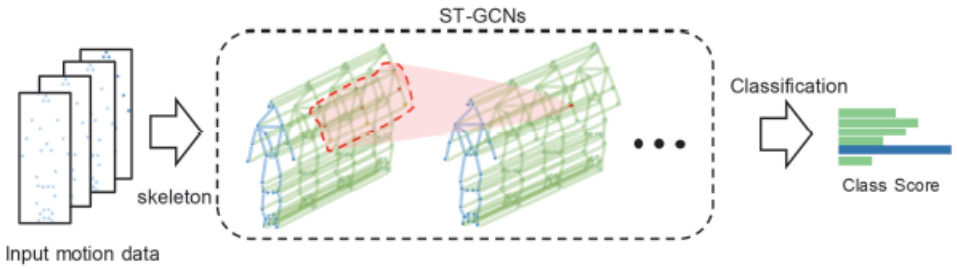

Figure 5. Graph Convolution [5]

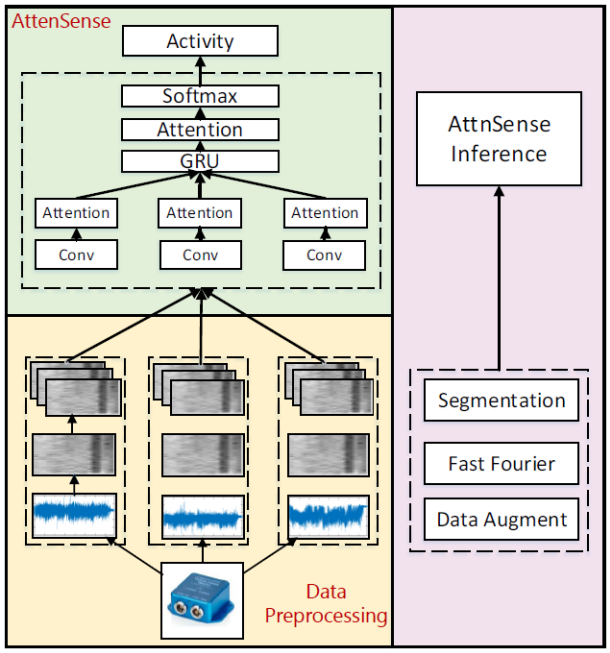

Figure 6. Overview of AttSense [6]

with binary weights, specifically, weights were binarized at each layer. The next forward method was used for the non-LSTM layer, with binary weights and scaling factors, and with real-values for the LSTM layer. The gradients were then computed using a backward algorithm. Finally, the learning rates and network parameters were updated. For the output layer, the Softmax function was used to generate responses ranging from 0 to 1. The posterior probability of the input sequence belonging to a certain activity can be seen as these outputs.
Experiment setup and Used tool: The proposed model is validated on PAMAP2 and Opportunity datasets. However, The tool used for the implementation is not mentioned this proposed article.
Comment: The suggested model gives 90.0% accuracy for 12 for PAMAP2 dataset as compared to other models and datasets.
Limitation: (i) The binarization strategy could be used to derive a relationship between model architecture properties and data size. (ii) problem caused by existing dataset distortions and improving sensor data calibration (iii) investigating the feature's impact on a more generalized system (iv) improving overall system efficiency by combining the advantages of B-BLSTM-RNNS and combination.

DCNN on Multichannel Time series for HAR [8]: Jian et al. represented this approach. In this approach, inertial sensors worn on the body are used to collect multichannel time series data which are multichannel time series data input and outputs are human activities which are predefined.

Functionality of Model: The CNN architecture is divided into five sections, as shown in figure.7. Each section in the initial two sections were made up of (a) a convolution layer for convolving the output of the preceding layer or the input and (b) a convolution layer for convoluting the output of the previous layer. (c) a ReLu layer for mapping the preceding layer's output, (d) a layer of max pooling for determining the highest feature map, and (e) to adjust the values of various feature maps a normalization layer. The third section contained a convolution, a ReLU, and a normalization layers; this section generates unified feature maps. The fifth layer was a network layer that was fully connected and converted the features into output classes.
Experiment setup and Used tool:The proposed approach is implemented on non-optimized MATLAB and validated on Opportunity and Hand gesture datasets.
Comment: The suggested model gives 94.1% accuracy for 12 activities for Hand Gesture dataset as compared to other models and datasets.
Limitation: When the CNN is trained and tested in parallel, the training and testing time can be significantly reduced.

An inertial accelerometer-based HAR [9]: Shaohua et al. represented this approach for Human Activity Recognition. They utilized CNN model in the architecture.
Functionality of Model: As illustrated in figure.8, the CNN architecture consisted of one input, one output, a fully connected layer, three convolutional layers, and three pooling layers. The data was received and preprocessed using the input layer. The data that was used was time series data. The features were extracted from the data using the convolution layer, and the number of features was reduced using the max pooling layer. The final layer before the output layer was used to combine the previous layers' results in order to calculate the score for each class. The output layer generated the results of the fully connected layers and outputs.
Experiment setup and Used tool:The proposed approach is implemented using python and for splitting the dataset, sklearn is used. The approach is validated on UCI and PAMAP2 datasets.
Comment: The suggested model gives 93.21% accuracy for 6 activities for UCI dataset as compared to other models and datasets.
Limitation: The structure of the neural network models could be further optimized, and a more detailed comparison could be performed.

DRNN-based activity recognition [10]: Masaya et al. demonstrated a high-throughput DRNN-based activity recognition approach. This method made use of LSTM.
Functionality of Model: A smartphone's three-axis acceleration provided direct input to the architecture. The DRNN was designed in such a way that each time's 3-axis acceleration data and 6 activity classes correlated to a 3-D input layer and a 6-D output layer, respectively. The LSTM unit was used for each internal layer. The error functions and activation were implemented using a cross entropy function and a Softmax function, respectively. During training, the weight was updated using the truncated BPTT gradient descent method. The quantity of internal layers, the maximum gradient, the truncated time, and the dropout probability were all made variable in order to find the best value. Finally, the network provided an output class for the activity.
Experiment setup and Used tool:The proposed approach is implemented using python and validated on HACS corpus dataset.
Comment: The suggested model gives 95.03% accuracy for 6 activities for HASC corpus dataset as compared to other models and datasets.
Limitation: The sequence data rate has been reduced. In this case, a post-processing method like HMM can be used.

DRNN for HAR [11]: Abdulmajid et al. represented a DRNN approach for identifying human activity based on LSTM.
Functionality of Model: obtained from the sensors were fed to the architecture. It was then segmented into T-dimensional windows and fed into a DRNN base in LSTM. The model generates class prediction scores for every time stamp. Then, these prediction scores are merged using late-fusion, which are then and fed into the Softmax function, which predicts the class of the activity.
Experiment setup and Used tool:The proposed approach is

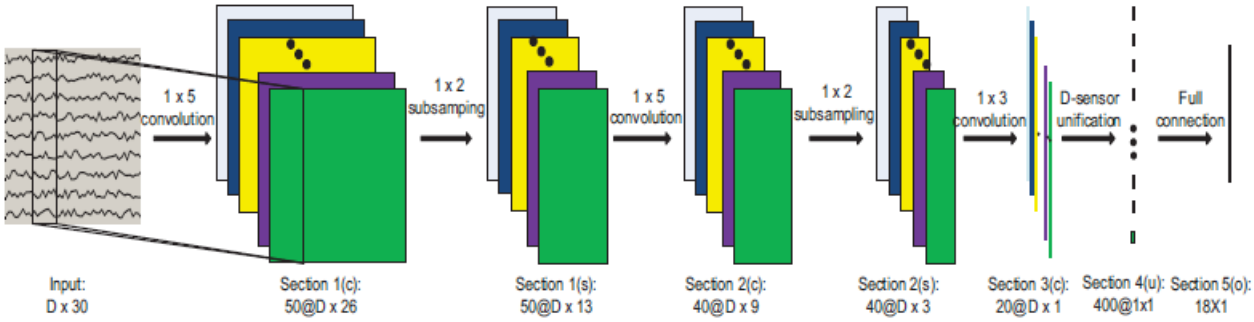

Figure 7. Architecture used for a multi sensor-based HAR problems [8]

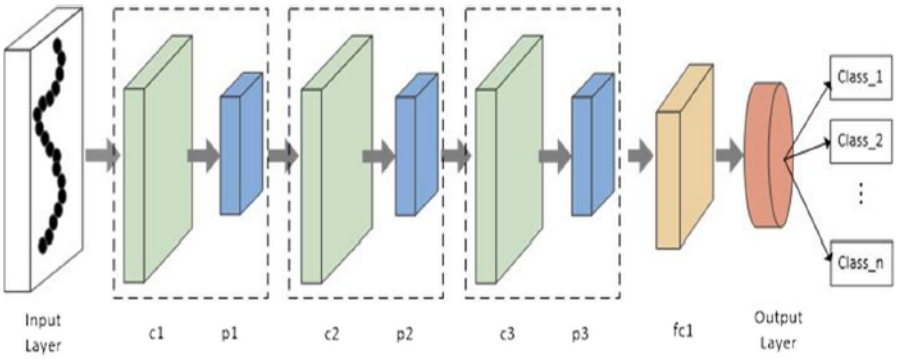

Figure 8. Example CNN architecture used for a smartphone inertial accelerometer-based architecture for HAR [9]

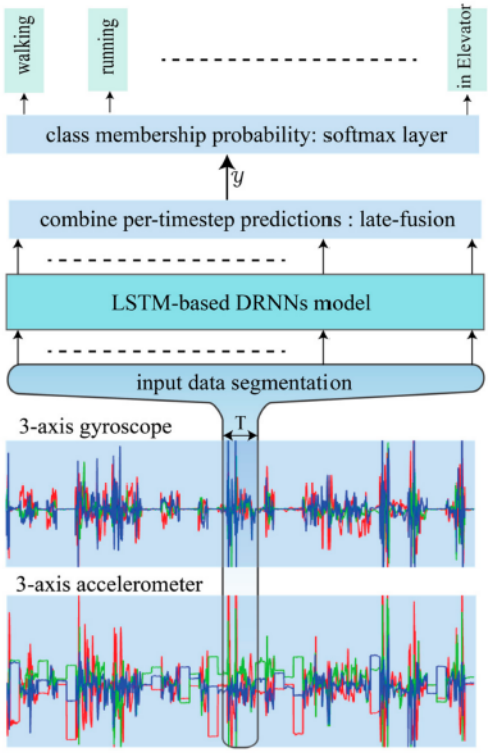

Figure 9. The architecture for DRNN [11]

implemented using GPU-based Tensorflow in python and validated on UCI-HAD, USC-HAD, Opportunity, Dalphnet FOG and skoda datasets.
Comment: The suggested model gives 97.8% accuracy for 11 activities for USC-HAD dataset as compared to other models and datasets.
Limitation: The model is not tested on large-scale and complex human activities. Transfer learning between different datasets could be investigated. Investigating resource-efficient DRNN implementation for low-power devices is another challenge.

Real-time driver activity recognition [12]: Chaopeng et al. represented this model for driver activity recognition utilizing ST-graph convolution LSTM.
Functionality of Model: Initially, the positions of the diver's upper body joins were recorded. The sliding window was used to smooth skeleton data using the temporal exponential mean filter. The driver skeleton graph was constructed using five layers of the GCN, as well as the representation of the spatial structure among joints. The spatial features of the GCN were transferred to a single layer of LSTM networks enhanced by attention mechanism. It extracted the temporal features from the sequence of frames. For balancing the loss value between other tasks related to driving and normal driving, the focal loss function was used. The final layer forecasted the driver's behavior.
Experiment setup and Used tool:The proposed approach is implemented using Tensorflow 1.4 in python 3.5 and validated on Self-made dataset.
Comment: The suggested model gives 88.8% accuracy for 8 activities (picking up objects, answering the phone, texting, using media, left or right checking, drinking, normal driving,) for self-made dataset as compared to other models.
Limitation: More driver activities could be taken into consideration in naturalistic driving conditions. In this approach, head motion and facial features were not taken into account. The approach is not considering the passengers in the vehicle.

GRU-based attention mechanism for HAR [13]: Nazmul et al. represented GRU-base attention mechanism for recognizing human activities. They have utilized the concept of the attention mechanism.
Functionality of model: Figure 11 depicts the architecture of this model which emphasis on the recurrent layers' hidden state outputs. For learning more complex features from the data captured by the sensors, the stacked layers are used. Prior to feeding the densely connected layers, the scores from attention mechanism from both of the layers were combined to form a hierarchy vector. Separately, the context-based and simplified attention scores were computed. Before and after applying attention and concatenating the attention scores, batch normalization was used. Three completely connected layers were used after the attention module. The first two layers were utilized in learning weights for various features extracted from the attention module. These layers were activated using ReLU. Dropout was used for regularization. Finally, the Softmax activation function was used to classify human activities at the final level.
Experiment setup and Used tool:The proposed approach is validated on Benchmark HAR dataset. However, The tool used for the implementation is not mentioned this proposed article.
Comment: The suggested model gives 94.16% accuracy for 6 activities for benchmark HAR dataset as compared to other models.
Limitation: By creating embedding from temporal data, more parallelizable models could be developed.

Human Activities of Daily living Recognition with GCN [14]: Nutchanun et al. proposed Human activities of Daily living recognition with Graph Convolutional Network(GCN).
Functionality of model: The architecture is divided into four steps, as shown in figure.12. First, data from image collections was gathered and chosen. The visual features were extracted first, followed by information about the word embedding vectors that represented the tagged annotation. The features were then graphed and analyzed. Finally, the data was used to train and validate the model, which was built using the selected features.
Experiment setup and Used tool:The proposed approach is validated on PASCAL VOC and LabelMe datasets. However, The tool used for the implementation is not mentioned this proposed article.
Comment: The suggested model gives 79.34% accuracy for 10 activities (Socializing, Watching TV, Relaxing, Working, Shopping, Eating, Housework, Commuting, Taking care of, and Preparing food) for Label Me dataset as compared to other models and datasets.
Limitation: The Wor2vec tool could not be able to handle the unknown words.

1D CNN-based HAR [15]: Song-Mi et al. represented one dimensional CNN-based human activity recognition.
Functionality of model: The acceleration signals x, y, and z are gathered and transmuted into vector magnitude data, as shown in fig.13. A 1D CNN was built using the vector magnitude data. The activities were classified using a 1D CNN. An input vector created from a fixed time-length accelerometer data was used as the input source. The convolution operation was then carried out with three different window sizes of 3, 4, and 5 and a stride size of one for all. The largest feature value was chosen using a max-pooling layer. To avoid overfitting, the resulting features were fed into a dropout layer. The final activity was predicted using the Softmax layer of a fully-connected network.
Experiment setup and Used tool:The proposed approach is implemented using Tensorflow in python and validated on WISDM dataset.
Comment: The suggested model gives 92.71% accuracy for 3 activities (walking, running, and staying still) for self-made dataset as compared to other models. *Limitation*: The walk activity signal contains ambiguous signals that are

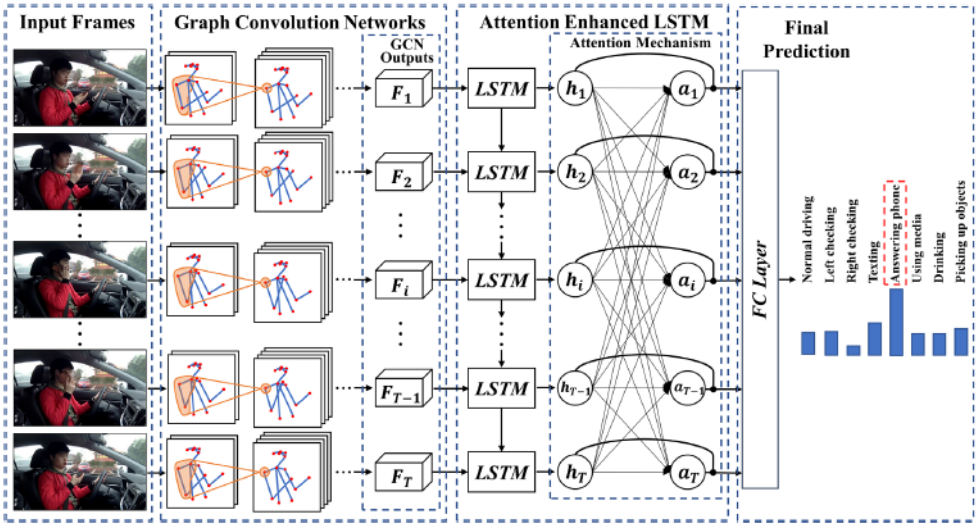

Figure 10. The architecture of Real-time driver activity recognition [12]

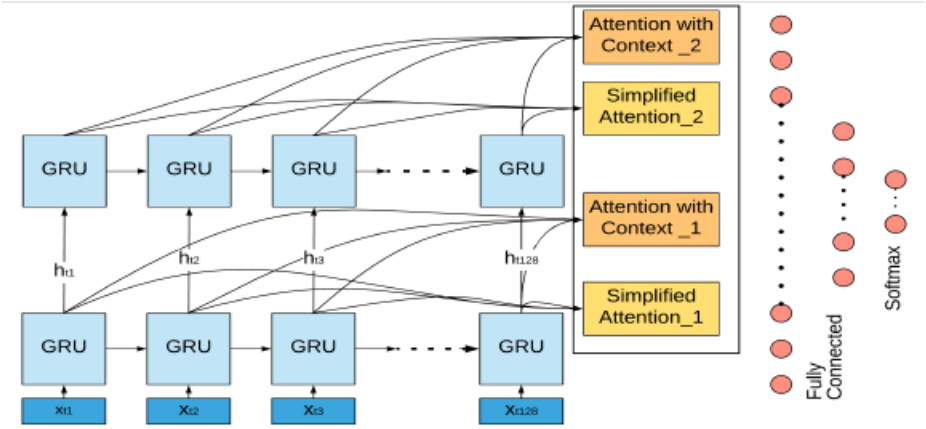

Figure 11. Architecture of GRU-based attention mechanism [13]

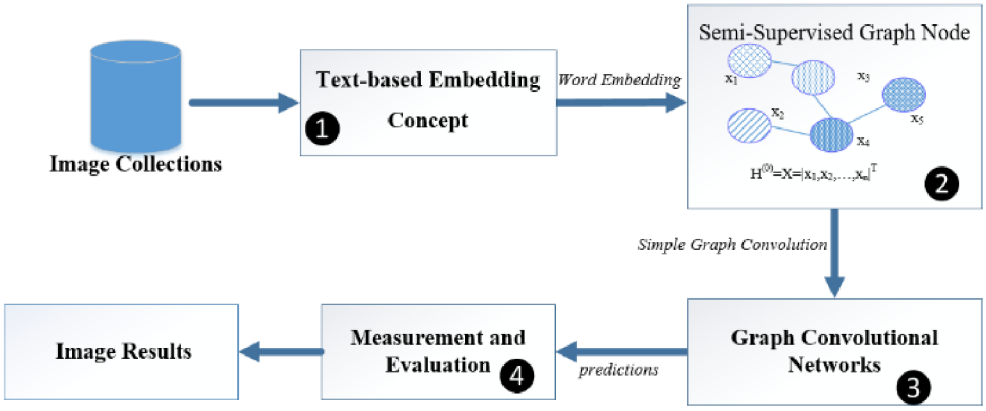

Figure 12. Overview of Human Activities of Daily living recognition framework [14]

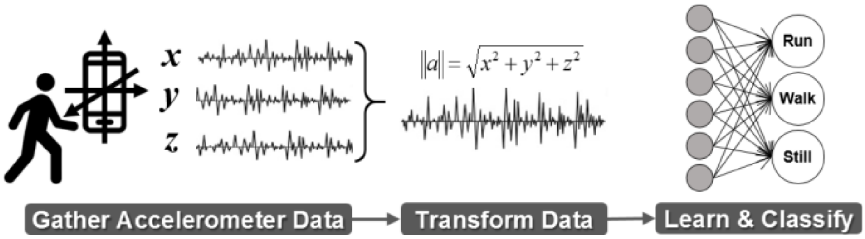

Figure 13. Overview of 1D CNN-based HAR [15]

interpreted as run and still, resulting in low precision in identifying walk activity.

DCNN for HAR [16]: Wenchao et al. represented the DCNN for HAR approach to Identify human activities. The model focuses on accuracy as well as reducing computational cost.
Functionality of model: from the gyroscope and accelerometer into an activity image that included the hidden relationships between any two signals. The activity image was fed into the DCNN, as shown in figure.14. The DCNN was made up of two convolution layers and two subsampling layers. The subsampling layers applied mean-pooling to the output of each convolution layer. The final layer is the fully-connected layer. To forecast the activity, the Softmax was used.
Experiment setup and Used tool:The proposed approach is implemented using MATLAB 2014 and validated on UCI, USC and SHO datasets.
Comment: The suggested model gives 99.93% accuracy for 3 activities (walking, sitting and laying) for SHO dataset as compared to other models and datasets.
Limitation: For very shallow neural networks, high-level features were not able to be learned. In the deep network, useful features might be filtered out during the convolution and subsampling.

HAR using Deep Learning with inertial sensors [17]: Tahmina et al. represented this approach. Convolution Neural Networks (CNN) architecture was utilized to identify human activities.
Functionality of model: To classify the activities more accurately, sensors were placed at five different locations on the lower body. The data was then preprocessed and segmented before being fed into a 6-channel 1D CNN network. To reduce the features, the CNN output was passed through the Max-pooling layer. It will then feed into the Softmax classifier layer and ReLu to classify the activity.
Experiment setup and Used tool:The proposed approach is implemented using Machine Learning Toolbox in MATLAB and validated on Self-made dataset.
Comment: The suggested model gives 97.01% accuracy for 6 activities for self-made dataset as compared to other models.
Limitation: The model could be tested on other activities also to verify the effectiveness of the model.

InnoHAR: A DNN for Complex HAR [18]: Cheng et al. represented InnoHAR for identifying human activities. This model is based on Recurrent neural network inception.

Functionality of Model: First, as shown in figure.16, the input data was routed through four Inception-like modules. Figure 17 depicts the structure of the Inception-like module. For eliminating the error caused by noise, after passing through two Inception-like modules, it was connected to the pooling layer. Finally, the output was fed into two GRU layers in order to extract the sequential temporal dependencies.
Experiment setup and Used tool:The proposed approach is implemented using Keras 2 in Python with Tensorflow as back-end and validated on Opportunity, PAMAP2 and Smartphones datasets.
Comment: The suggested model gives 94.6% accuracy for 18 activities for Opportunity dataset as compared to other models and datasets.
Limitation: The network structure could be more adjusted, including the connection method and the size of kernels.

LSTM-CNN for HAR [19]: Kun et al. represented LSTM-CNN architecture for human activity recognition. This approach combines convolution layers with LSTM.
Functionality of Model: The preprocessed data was initially fed into a two-layer LSTM for temporal feature extraction. The spatial features were then extracted using two convolution layers. To reduce overfitting, a maxpooling layer was placed between these two convolution layers. To reduce global modal parameters, a global average pooling layer was used after the convolution layer. Following the GAP, batch normalization (BN) was used to normalize and reconstruct the input data on training sample of each batch, ensuring the previous layer's stability and improving training speed and accuracy. Finally, the output of the Softmax classifier was obtained.
Experiment setup and Used tool:The proposed approach is implemented using Keras in Python with Tensorflow as back-end and validated on Opportunity, WISDM and UCI datasets.
Comment: The suggested model gives 95.85% accuracy for 6 activities (Downstairs, Jogging, Sitting, Standing, Upstairs, Walking) for WISDM dataset as compared to other models and datasets.
Limitation: The GAP layer focuses on the model's training pressure on the convolution layer, causing the model to

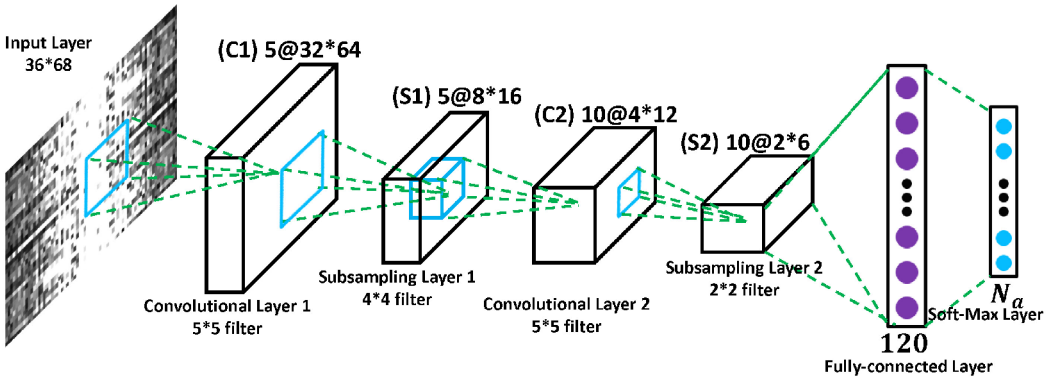

Figure 14. Architecture of DCNN [16]

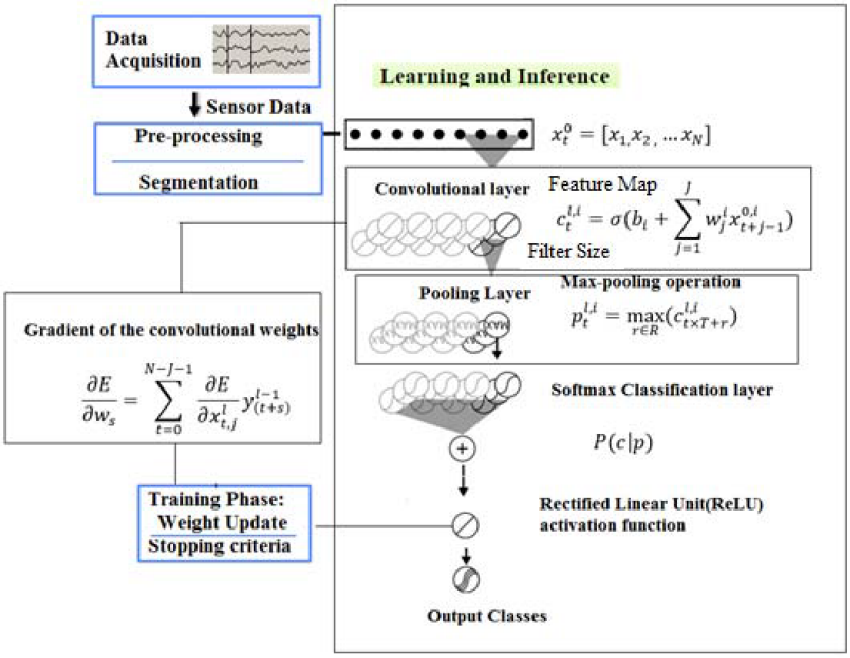

Figure 15. Architecture of DCNN [17]

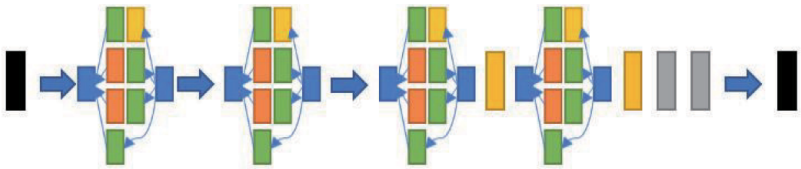

Figure 16. Overall architecture of InnoHAR [18]

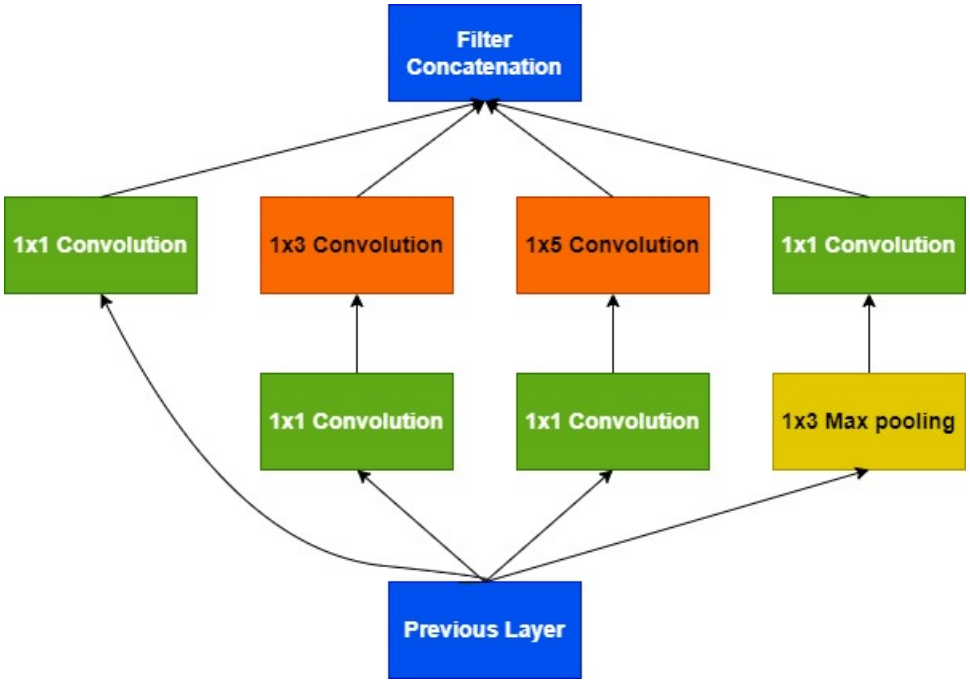

Figure 17. Example: Inception-like module in InnoHAR [18]

converge slowly.

Semi-supervised deep learning using DLSTM [20]: Quingchang et al. represented a semi-supervised deep learning using DLSTM for HAR. Labelled and unlabeled data were used as an input source of the architecture.
Functionality of Model: For increasing the training dataset and some processes, an augmentation technique was used, as shown in figure.18. The low-level features were derived from the simple statistics. The high-level features were then taught to the DLSTM network. To improve the generalization of the deep architecture, the dropout method was then used as a regularizer. The supervised loss was then calculated with some labels, and the unsupervised loss was calculated by comparing the unlabeled prediction to the ensembled predictions in previous epochs. The results were classified by the Softmax dense layer, which was the final layer.
Experiment setup and Used tool:The proposed approach is implemented using Scikit Learn in Python and validated on UCI dataset.
Comment: The suggested model gives 92.1% accuracy for 6 activities for UCI dataset as compared to other models.
Limitation: The recognition of the unseen classes is not possible with this architecture.

A Lightweight DL model for HAR [21]: Preeti et al. proposed a lightweight DL model for HAR. This model necessitates less computational power which makes appropriate it suitable for deployment over edge devices.
Functionality of Model: The input source of the model was the fixed-size window accelerometer reading. The segmented readings were fed into the Lightweight RNN-LSTM model, which was made up of two hidden layers, each with 30 neurons. In this model, which employed a set of rules, the Softmax classifier was used to merge output from various states into a single final output.
Experiment setup and Used tool:The proposed approach is implemented using python and Tensorflow and validated on WISDM dataset.
Comment: The suggested model gives 95.78% accuracy for 6 activities for UCI dataset as compared to other models.
Limitation: The model could be tested on complex activities before being deployed. It was used only on Raspberry Pi3.

HAR from accelerometer data using CNN [22]: Ignatov represented this approach using convolution neural networks.
Functionality of Model: The input was fed into the Convolution layer. A non-linear activation function follows this layer to learn non-linear decision boundaries, as shown in figure.19. To reduce and summarize the obtained representation, a pulling layer was added after the convolution layer. After several convolution and max-pooling layers, the output layer was flattened into a 1D vector and used for classification. CNN has one or more fully-connected layers on top to learn non-linear classification.. Finally, the last layer's output was fed into a Softmax layer, which computed the probability distribution over all the predicted classes.
Experiment setup and Used tool:The proposed approach is implemented using Tensorflow and validated on WISDM and UCI datasets.
Comment: The suggested model gives 97.63% accuracy for 6 activities for UCI dataset as compared to other models

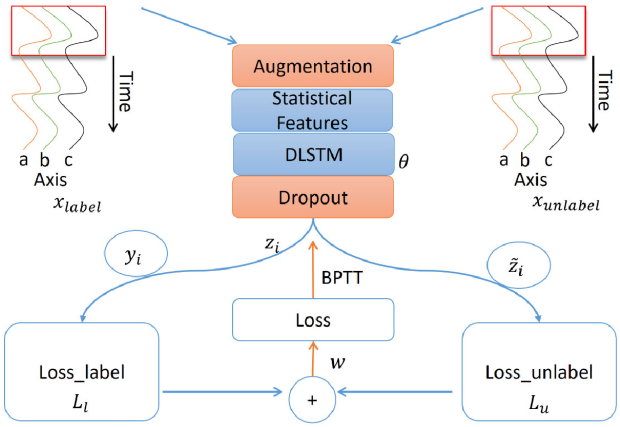

Figure 18. Structure of the temporal ensambling of DLSTM [20]

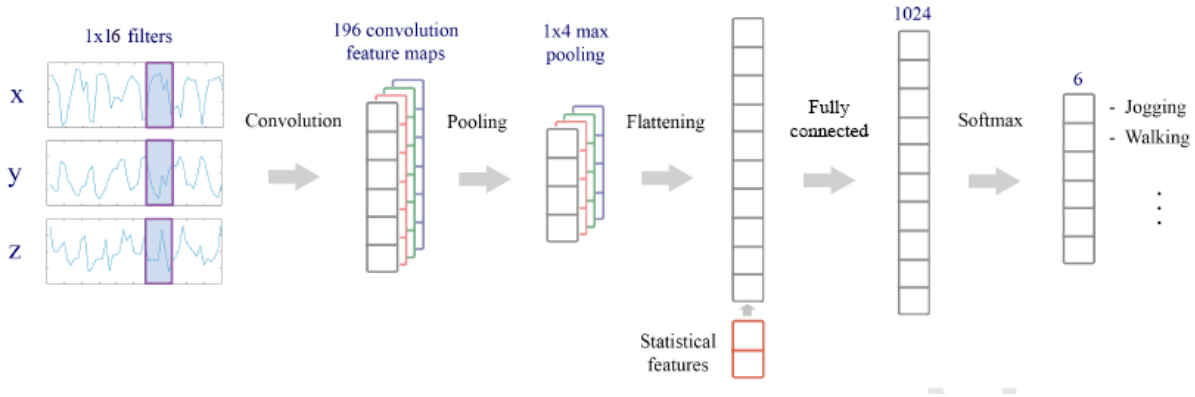

Figure 19. Example The architecture of a user-independent deep learning approach for online HAR classification [22]

and datasets.
Limitation: The model architecture could be tested on other publicly available datasets to test the efficiency over them.

CnvLSTM-FC model for HAR [23]: Jian et al. represented a hybrid deep architecture depending on ELM classifier, LSTM recurrent units, and convolution operations.
Functionality of Model: The CNN architecture is not well suited to dealing with time series data. To deal with time series data, LSTM was introduced for modeling temporal dependencies on output features of CNN layer, as shown in figure.20. In the following step, ELM classified was used to classify the features that contained temporal information, which classified the activity class.
Experiment setup and Used tool:The proposed approach is implemented using python 2.7.11 and validated on Opportunity dataset.
Comment: The suggested model gives 91.8% accuracy for 18 activities for Opportunity dataset as compared to other models.
Limitation: The sequential learning adaptive capability of the model could be improved; transfer learning approach could be utilized.

HAR using BMI and DL [24]: Dobhal et al. represented an approach to recognize HAR using BMI and DL.
Functionality of Model: As shown in figure. 21, in this approach, first of all, from each frame, background is subtracted for obtaining only foreground, in this case a person, using GMM. Furthermore, the Binary Motion Image (BMI) is calculated which represents a sequence of video frames in a single frame. The same will be given as the input to the CNN which will classify the activity.
Experiment setup and Used tool:The proposed approach is implemented using MATLAB and validated on MSR Action 3D dataset.
Comment: The suggested model gives 98.5% accuracy for 20 activities for MSR Action 3D dataset.
Limitation: This approach have minor level of invariance to scale changes, rotation and translation.

Sensor-Based HAR [25]: Nafea et al. represented Sensor-based HAR using CNN and BiLSTM.
Functionality of Model: In this model, CNN and BiLSTM are used to form two-stream DL architecture. The activity recognition is taking place through Features Fusion obtained from two-stream. BiLSTM understands the spatial deep-

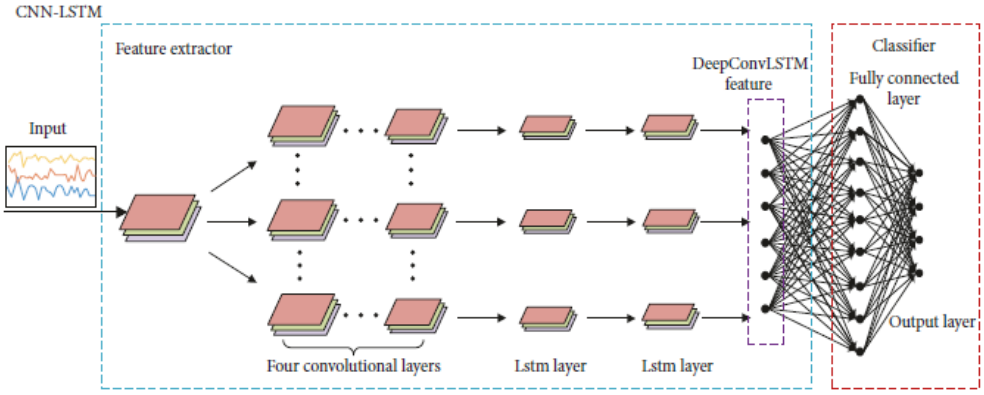

Figure 20. Example of The architecture for CnvLSTM-FC model for HAR [23]

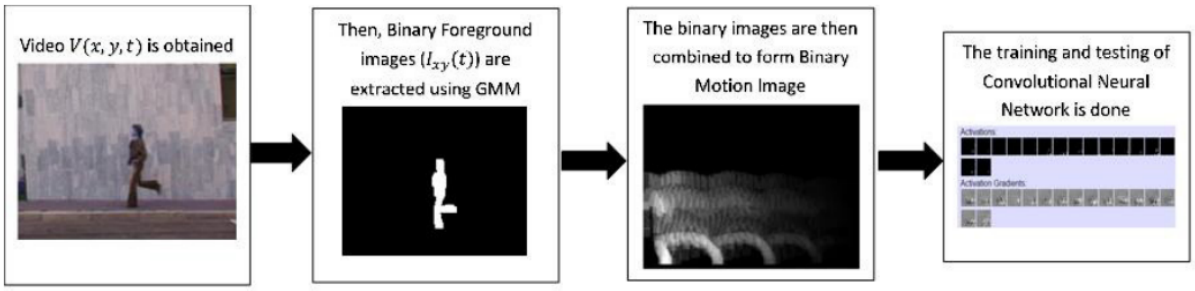

Figure 21. Architecture of HAR using BMI and DL [24]

learning map's underlying temporal relationship while CNN with many convolution layers and different kernel dimensions obtains spatial relationship.
Experiment setup and Used tool:The proposed approach is implemented using Tensorflow, Keras and Theano in python and validated on UCI and WISDM datasets.
Comment: The suggested model gives 98.53% accuracy for 6 activities for WISDM dataset and 97.05% accuracy for 6 activities for UCI-HAR dataset.
Limitation: The features from CNN and BiLSTM could be derived through automatic feature extraction.

Multi-input CNN-GRU [26]: Dua et al. proposed Multi-input CNN-GRU approach to identify human activities.
Functionality of the model: This model comprises of three heads. Identical inputs are fed to each head. Two subsequential Conv1D layers make up each head with different filter sizes for capturing different temporal local dependencies. This Conv1D is applied to the input vector. Conv1D uses ReLu as the activation function. After the second Conv1D, a dropout layer is placed. A Max-pooling layer follows the dropout layer. Next, the flatten layer flatten the feature vector to make it suitable for GRU layer. Two GRU layers are utilized here. The dense layer now receives the combined output from all three heads. After that, batch normalization is done to the Dense layer's output. A Softmax function is following the Batch normalization for final classification of the activity.
Experiment setup and Used tool:The proposed approach is implemented using Keras in python with Tensorflow as back-end and validated on UCI-HAR, Opportunity and WISDM datasets.
Comment: The suggested model gives 97.21% accuracy for 6 activities for WISDM dataset, 96.20% accuracy for 6 activities for UCI-HAR dataset and 95.27% accuracy for 12 activities for PAMAP2 dataset.
Limitation: The model could be applied to other state-of-art datasets.

DNN for multi-view HAR [27]: Putra et al. represented a DNN for multi-view HAR.
Functionality of the model: input to this model are the images taken from the multiple cameras from multiple angles for resolving occlusion problem. These images are fed to a pre-trained CNN for capturing spatial information. Author assumes that significant transformation is taking place in only certain part of the image. For that purpose, features are fed to attention layer to give weight to the important features. To get the temporal data, LSTMRes layer is utilized. Finally, for combining prediction scores from the Softmax, this model uses arithmetic mean or geometric mean.
Experiment setup and Used tool:The tool for implemen-

tation is not mentioned in the proposed article. However, the proposed approach is validated on IXMAS and i3DPost datasets.
Comment: The suggested model gives 96.37% accuracy for 13 activities for IXMAS dataset and 96.87% accuracy for 10 activities for i3DPost dataset.
Limitation: (i) This model does not perform well for ambiguous actions. (ii) The models could be tested on other state-of-art datasets. (iii) This model observes only observes self-occlusion in the datasets.

ConvAE-LSTM [28]: Thakur et al. represented Convolution Auto encoder-LSTM approach.
Functionality of the model: The model consists of three modules. First module is a ConvAE, which is composed of a convolution layer, pooling layer and deconvolution layer. The output of ConvAE must be compatible with the input format of LSTM, for that purpose, the flatten layer is used next. The output of the LSTM is fed to the fully connected layer. It will give the high-level representation. Finally, a Softmax layer is used to classify the physical activity.
Experiment setup and Used tool:The proposed approach is implemented using Keras in python with Tensorflow as back-end and validated on UCI-HAR and WISDM datasets.
Comment: The suggested model gives 98.14% accuracy for 6 activities for UCI-HAR dataset and 98.67% accuracy for 6 activities for WISDM dataset.
Limitation: (i) The model could be compared with other recent DL models and also be tested on other datasets. (ii) The applicability in real-life application should be analyzed.

HAC using 3DCNN [29]: Vrskova et al. proposed classification of Human activities using 3DCNN.
Functionality of the model: This model utilizes 3DCNN layers for improving the identification of moving images. Each layer in 3DCNN contains 3D filter. This layer creates a convolution map. A Max-pooing layer follows 3DCNN which is used to reduce the image size. Next, the Batch normalization is utilized to normalized each batch's previous layer. Next, the dense layer and flatten layer are used to forma the output vector.
Experiment setup and Used tool:The proposed approach is implemented using Keras in python with Tensorflow as back-end and validated on UCF YouTube Action dataset, modified UCF101 dataset, and full UCF101 dataset.
Comment: The suggested model gives 74.20% accuracy for 11 activities for UCF YouTube Action dataset, 84.40% accuracy for 15 activities for modified UCF101 dataset and 79.90% accuracy for 15 activities for full UCF101 dataset.
Limitation: Identification of human non-standard behavior is difficult.

3. Datasets

Datasets are one of the important components of Human Activity Recognition Algorithms. They are used as an input source to the algorithm. Some of the commonly used datasets for HAR are explained briefly as below:

A. iSPL

The iSPL dataset [**?**] is a three-activity dataset containing gyroscope and triaxial raw accelerometer signals collected from a WithRobotTM sensor. The sensor was attached to a subject's left wrist. Four subjects ranging in age from 25 to 40 years were used. The iSPL consists of 1590 samples, which are then randomly divided into 1272 train samples and 318 test samples. Each data sample includes 128 sensor readings for each of the nine signal types.

B. UCI HAR

The UCI HAR dataset [30] is a 6-activity (Laying, Standing, Sitting, Walking Downstairs, Walking Upstairs, Walking) dataset containing 3D (x,t,z) raw signals collected from a smartphone's gyroscope and accelerometer. A smartphone was strapped to a subject's waist. The experiment was carried out on 30 subjects ranging in age from 19 to 48 years. Each subject completed the activities mentioned before. The dataset contains 7352 train samples and 2947 test samples. To preprocess the sensor signals, noise filters and a sampled fixed-width sliding window were used.

C. Opportunity dataset

The Opportunity dataset [31] is an 18-activity dataset where on-body sensors were used to record human activities. The recording sessions of two kinds were carried out with the help of four participants. First, the subject completed a predefined set of activities 20 times during a Drill session. Second, in the Daily living activity session, participants completed high-level tasks with greater flexibility in the order of activities. The dataset also includes a Null activity, which specifies the period with no activity or activity that is irrelevant.

D. PAMAP2

The PAMAP2 [32] is an 18-activity physical activity monitoring dataset. The activities were carried out by 9 people. Three inertial measurement units recorded the activities. These units are a magnetometer, an accelerometer, and a gyroscope, , and, which are located on the ankle, chest, and hand respectively.

E. WISDM

The WISDM [33] is a 6-activity dataset (Downstairs, Jogging, Sitting, Standing, Upstairs, Walking) that includes activities recorded with an Android phone in the subject's front pocket. This experiment used a total of 36 subjects. The activity was recorded using an accelerometer with a sampling frequency of 20Hz. This dataset contains 1098209 samples. The WISDM dataset is unbalanced because the activities in it have an uneven number of samples.

F. MSRC-12

The MSRC-12 [4] dataset contains a sequence of human activities (kick, beat both, change weapon, had enough, throw, wind it up, bow, shoot, goggles, push right, duck, and lift arms). Thirty subjects participated in the activities,

which were recorded with a depth camera. It has a total of 6244 activities. The human skeletal joins were included in the dataset.

G. Skoda

The Skoda [34] dataset is a 10-activity dataset that describes the activities, in a car production domain, carried out by assembly-line workers. Numerous accelerometers were placed on both the hands of each worker during the activities. The dataset also includes a Null activity, which specifies the period with no activity.

H. Hand Gesture

The Hand Gesture dataset [35] is a 12-activity dataset containing various types of human hand movements. The activities were carried out by a total of two subjects. The subjects performed hand movements in daily life with eight different hand gestures and three different gestures while playing tennis. It, like the Opportunity and Skoda datasets, has a null activity that specifies the period with no activity or non-relevant activity.

I. USC-HAD

The USC-HAD [36] is a 12-activity dataset contains basic human activities (running, walking downstairs, walking upstairs, walking right, walking left, walking forward, sitting, standing, jumping, sleeping, in elevator down, in elevator up). A 3D accelerometer and gyroscope sensor were attached to the subject's front hips to collect the data.

J. Benchmark HAR dataset

The Benchmark HAR dataset [37] is a 6-activity dataset containing time series data for these activities (laying, standing, sitting, walking downstairs, walking upstairs, walking). The accelerometer and gyroscope sensors were used to collect this data. It divided the data into 30 percent for testing and 70 percent for training at random. To preprocess the sensor signals, noise filters and a sampled fixed-fixed width sliding window were used. Training data for 7352 samples was generated by 21 subjects, and testing data for 2947 samples was generated by 9 subjects.

K. Heterogeneous dataset

The Heterogeneous dataset [38] is a 6-activity (biking, walking, sitting, standing, climb-up and climb-down) dataset containing gyroscope and accelerometer sensing data. The activities were accomplished by a total of 9 subjects. The activities were carried out using a variety of smartphones and smart watches. This increases the task's complexity and aids in testing the model's robustness.

4. Evaluation Parameter

For the purpose of comparative analysis, we have selected accuracy as the evaluation parameter. Accuracy is a metric which is used to evaluate classification models. Accuracy can be stated as :

Accuracy = # of correct predictions / total # of predictions [39]

According to the equation, the accuracy can be stated as the correct number of predictions out of the total number of predictions.

5. Comparative Analysis of HAR Models

The comparative analysis of HAR using deep neural networks is done on five different parameters. The HAR can be Vision based or Sensor based. The Vision based techniques are dependent on cameras whereas the Sensor based techniques are dependent on sensors which may be put on the human body. The comparison can also be carried out based on the dataset used and the accuracy achieved by the model on different activities.

Table 1 represents 28 recent HAR models based on deep neural networks. Twenty one models are sensor based and the remaining models are Vision based. In sensor-based models, seven models are using Convolution Neural Network architecture, eight models are using Recurrent Neural Network (RNN) architecture, six models are using combination of CNN and RNN architectures and one model is using combination of CNN, RNN with the attention mechanism. On other side, in vision-based models, two models are using CNN architecture, one model is using RNN architecture, one model is using Graph Convolution Neural Network (GCN) architecture, one models is using combination of GCN and RNN, one model is using GCN with the attention mechanism and one model is using combination of CNN, LSTM and attention mechanism.

Various datasets are used as an input source of the model. The most commonly used datasets are Opportunity dataset, PAMAP2, UCI and WISDM. The comparative analysis of the average accuracy of identifying different activities by different models on these commonly used datasets is shown in figure 22, 23, 24 and 25, respectively.

As shown in Figure 22, five models have used the Opportunity dataset as an input source. Among these models, the model used in [18] has the highest predictive accuracy of 94.6% on the mentioned dataset on 18 activities. As shown in Figure 23, five models have used the PAMAP2 dataset as an input source and among these models, model used in [26] has the highest predictive accuracy of 95.27% on the mentioned dataset on 12 activities.As shown in Figure 24, six models have used the WISDM dataset as an input source and among these models, model used in [28] has the highest predictive accuracy of 98.67% on the mentioned dataset on 6 activities.As shown in Figure 25, ten models have used the UCI dataset as an input source and among these models, model used in [28] has the highest predictive accuracy of 97.63% on the mentioned dataset on 6 activities.

TABLE I. Comparative Analysis of HAR Models

Title and Year	Architecture	Sensor/Vision based	Dataset	No. of Activities Recognized	Accuracy (%)
CNN-LSTM (2020) [2]	CNN-LSTM	Sensor	iSPL	3	99.00
			UCI HAR	6	92.0
Acceleration-based HAR using CNN (2015) [3]	CNN	Sensor	Self-made dataset	8	93.80
RNN-based HAR (2016) [4]	RNN	Vision	MSRC-12	12	99.55
ST-GCN (2019) [5]	ST-GCN	Vision	Carecom nurse care activity dataset	7	57.00
AttnSense for multimodal human activity recognition (2019) [6]	Attention Mechanism-CNN-GRU	Sensor	Heterogeneous	6	96.50
			Skoda	10	93.10
			PAMAP2	12	89.30
Binarized-BLSTM-RNN based HAR (2016) [7]	Binarized-BLSTM-RNN	Sensor	PAMAP2	12	90.00
			Opportunity Activity Recognition Dataset	18	74.00
DCNN on Multichannel Time series for HAR (2015) [8]	CNN	Sensor	Opportunity Activity Recognition Dataset	18	87.70
			Hand Gesture Dataset	12	94.10
An inertial accelerometer-based HAR (2019) [9]	CNN	Sensor	UCI	6	93.21
			PAMAP2	9	91.00
DRNN-based activity recognition (2016) [10]	DRNN	Sensor	HASC corpus	6	95.03
DRNN for HAR (2017) [11]	DRNN	Sensor	UCI-HAD	6	96.70
			USC-HAD	11	97.80
			Opportunity Activity Recognition Dataset	18	92.00
			Daphnet FOG	2	93.00
			Skoda	11	92.60
Real-time driver activity recognition (2020) [12]	Spatialtemporal Graph Convolution LSTM networks with attention	Vision	Self-made dataset	8	88.80
GRU-based attention mechanism for HAR (2019) [13]	GRU-attention mechanism	Sensor	Benchmark HAR dataset	6	94.16
Human Activities of Daily living Recognition with GCN (2020) [14]	GCN	Vision	PASCAL VOC	10	78.67
			LabelMe	10	79.34
1D CNN-based HAR (2019) [15]	LSTM-RNN	Sensor	WISDM	6	94.00

DCNN for HAR (2015) [16]	DCNN	Sensor	UCI	3	95.18
			USC	3	97.01
			SHO	3	99.93
HAR Deep Learning with inertial sensors using (2017) [17]	CNN	Sensor	Self-made dataset	6	97.01
InnoHAR: A DNN for Complex HAR (2018) [18]	Inception NNRNN	Sensor	Opportunity Dataset	18	94.60
			PAMAP2	18	93.50
			Smartphones dataset	6	94.50
LSTM-CNN for HAR (2020) [19]	LSTM-CNN	Sensor	UCI	6	95.78
			WISDM	6	95.85
			Opportunity Activity Recognition Dataset	18	92.63
Semisupervised deep learning using DLSTM (2018) [20]	DLSTM	Sensor	UCI	6	92.10
A Lightweight DL model for HAR (2019) [21]	RNN-LSTM	Sensor	WISDM	6	95.78
HAR from accelerometer data using CNN (2017) [22]	CNN	Sensor	WISDM	6	93.32
			UCI	6	97.63
CnvLSTM-FC model for HAR (2018) [23]	CNN-LSTM-ELM	Sensor	Opportunity Dataset	18	91.80
HAR using BMI and DL (2015) [24]	CNN	Vision	MSR Action 3D	20	98.50
Sensor-Based HAR (2021) [25]	CNN-BiLSTM	Sensor	WISDM	6	98.53
			UCI-HAR	6	97.05
Multi-input CNN-GRU (2021) [26]	CNN-GRU	Sensor	UCI-HAR	6	96.20
			WISDM	6	97.21
			PAMAP2	12	95.27
DNN for multi-view HAR (2022)[27]	CNN-attention-LSTM	Vision	IXMAS	13	96.37
			i3DPost	12	96.87
ConvAE-LSTM (2022) [28]	ConvolutionAE-LSTM	Sensor	UCI-HAR	6	98.14
			WISDM	6	98.67
HAC using 3DCNN (2022) [29]	3DCNN	Vision	UCF YouTube Action dataset	11	74.20
			Full UCF101	15	79.90
			Modified UCF101	15	84.40

6. Applications

Some of the application of HAR are recapitulated and explained briefly as below:

Video analysis based on content: If the activities in a video are recognized, it will be easier to categorize the videos based on the contents. It can improve user experience, content storage, and summarization. It has the potential to be useful in video sharing and other applications [40].

Behavioral Biometrics: Behavioral biometrics is used to identify a person based on patterns in their behavior. When compared to traditional biometric identification methods,this method requires very little or no human intervention [41].

Security and surveillance: HAR applications can be used in security and surveillance systems. It will be easier to identify unusual activities if vision-based activity recognition is integrated into surveillance systems. This reduces the need to manually analyze multiple videos for the same purpose [40].

Interactive applications and environments: The primary goal of the Human Computer Interaction system is to comprehend human activities in order to respond to human activities. This type of system receives input from gestures or actions and responds to the gestures or actions. This type of system can help in the development of robots and computers that can effectively respond to and interact with humans [40].

7. Challenges

A. Vision-based HAR

Some of the challenges in vision-based HAR are listed and explained briefly as below:

Changes in illumination: Some video parameters, such as contrast, brightness, and so on, may change dynamically, or they may be affected by factors such as environmental changes, etc. [42], [43].

The shadow effect: A person's or an object's shadow may cause false detection or tracking of activity.

Human behavior: When humans perform multiple tasks at once, it becomes more difficult to identify the activity [40].

Intra-class variability: It occurs when different users perform the same activity differently [44].

Inter-class similarity: Characteristics of fundamentally different classes may be similar. Activities such as skipping and running are examples [45].

Occlusions, partial or total: The subject whose activity is to be identified may be occluded by another object, making it difficult to recognize the activity [40].

Bootstrapping: The training environment's background may differ from the real-world environment's background. [40].

Camera jitter: A low-resolution camera or low-quality recording device degrades video quality [44].

Self-occlusions: When one body part of an object occludes another, it becomes difficult to recognize the activity [40].

Scaling: The device that detects human activity can be placed close to or far away from the person performing the activity [40].

Camera automatic adjustments: The automatic adjustment feature in modern cameras may make identifying activity difficult because in different frames, the same image may appear differently [40].

B. Sensor-based HAR

Some of the challenges in vision-based HAR are listed and explained briefly as below:

Smartphone and wearable sensor location: The position of the smartphones must be taken into account when retrieving data because the gyroscope determines orientation and accelerometers handle axis-based motion sensing. Otherwise, this could lead to an incorrect interpretation of the activity [46].

Requirements of sensors: When retrieving data from users, the complexity of the HAR algorithm grows in direct proportion to the type of sensor used, the number of smartphones used, and the location of the smartphone [46].

Quantity of sensors to choose: Plethora of sensors are available in a smartphone from which to choose. Among these sensors, the sensors should be chosen precisely so that the process of identifying activity in real-time applications becomes easier and more convenient [46].

8. Conclusion

This paper represents a comprehensive analysis of Human Activity Recognition using Deep Learning. Challenges faced in HAR differ based on the approach used i.e. Vision-based or Sensor-based. The basic HAR approach includes Data collection, Data pre-processing, Feature extraction, Learning and Activity Recognition. The commonly used datasets for HAR are Opportunity dataset, WISDM, PAMAP2, UCI, etc. In the comparative analysis, we observed that in Vision-based approach, [4] model gives good performance 99.55% accuracy along with 12 activities with MSRC-12 dataset and in Sensor-based approach, [16] model gives good performance 99.93% accuracy along with 3 activities with SHO dataset. It is also observed that for Opportunity dataset, [18] model gives good performance with 94.6% accuracy along with 18 activities, for PAMAP2 dataset, [26] model gives good performance with 95.27% accuracy along with 12 activities, for WISDM dataset, [28] model gives good performance with 98.67% accuracy along with 6 activities, and for UCI-HAR dataset, [28] gives good performance with 98.14% accuracy along with 6 activities.

References

[1] J. Yang, J. Lee, and J. Choi, "Activity recognition based on rfid object usage for smart mobile devices," *Journal of Computer Science and Technology*, vol. 26, no. 2, pp. 239–246, 2011.

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

[3] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE international conference on systems, man, and cybernetics*. IEEE, 2015, pp. 1488–1492.

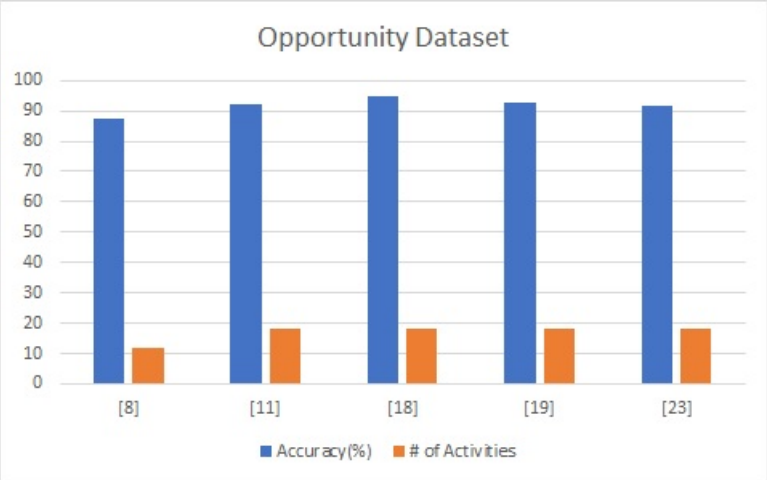

Figure 22. Comparative analysis of different models on Opportunity dataset

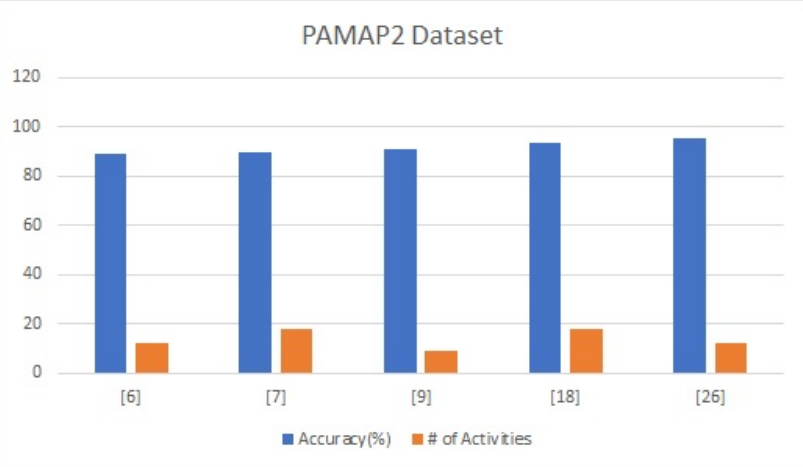

Figure 23. Comparative analysis of different models on PAMAP2 dataset

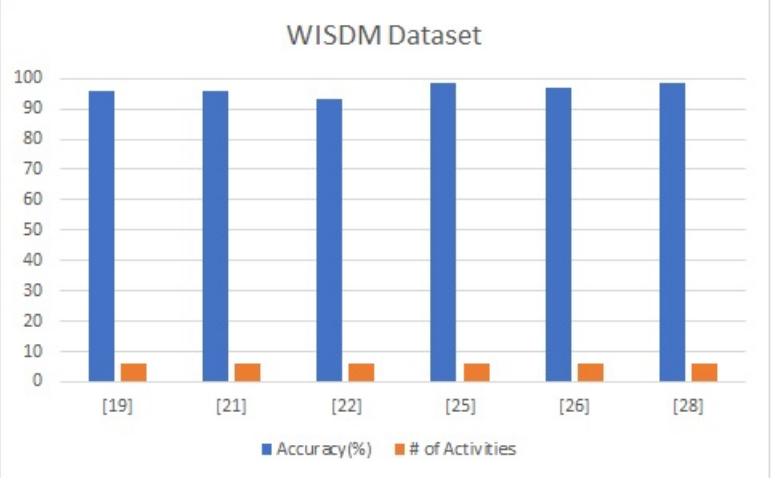

Figure 24. Comparative analysis of different models on WISDM dataset

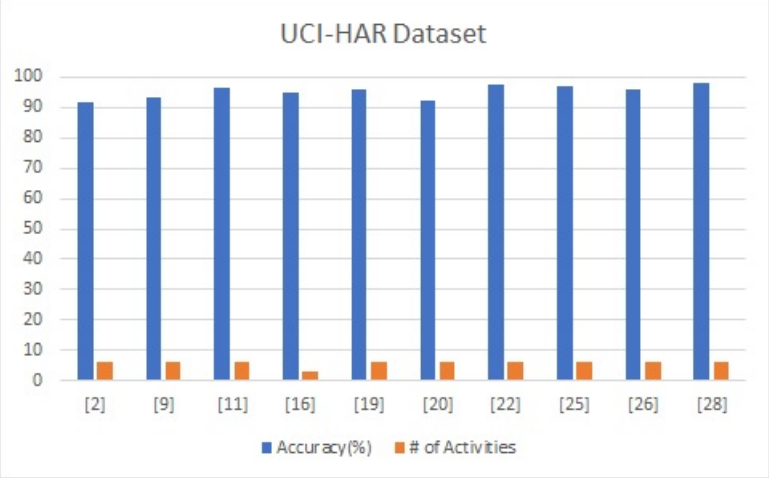

Figure 25. Comparative analysis of different models on UCI-HAR dataset

[4] S. Park, J. Park, M. Al-Masni, M. Al-Antari, M. Z. Uddin, and T.-S. Kim, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," *Procedia Computer Science*, vol. 100, pp. 78–84, 2016.

[5] X. Cao, W. Kudo, C. Ito, M. Shuzo, and E. Maeda, "Activity recognition using st-gcn with 3d motion data," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 689–692.

[6] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition." in *IJCAI*, 2019, pp. 3109–3115.

[7] M. Edel and E. Köppe, "Binarized-blstm-rnn based human activity recognition," in *2016 International conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2016, pp. 1–7.

[8] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[9] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 743–755, 2020.

[10] M. Inoue, S. Inoue, and T. Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput," *Artificial Life and Robotics*, vol. 23, no. 2, pp. 173–185, 2018.

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

[12] C. Pan, H. Cao, W. Zhang, X. Song, and M. Li, "Driver activity recognition using spatial-temporal graph convolutional lstm networks with attention mechanism," *IET Intelligent Transport Systems*, vol. 15, no. 2, pp. 297–307, 2021.

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

[14] N. Chinpanthana and Y. Liu, "Human activities of daily living recognition with graph convolutional network," in *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, 2020, pp. 305–310.

[15] S.-M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *2017 ieee international conference on big data and smart computing (bigcomp)*. IEEE, 2017, pp. 131–134.

[16] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.

[17] T. Zebin, P. J. Scully, and K. B. Ozanyan, "Human activity recognition with inertial sensors using a deep learning approach," in *2016 IEEE sensors*. IEEE, 2016, pp. 1–3.

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

[19] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[20] Q. Zhu, Z. Chen, and Y. C. Soh, "A novel semisupervised deep learning method for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3821–3830, 2018.

[21] P. Agarwal and M. Alam, "A lightweight deep learning model for human activity recognition on edge devices," *Procedia Computer Science*, vol. 167, pp. 2364–2373, 2020.

[22] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018.

[23] J. Sun, Y. Fu, S. Li, J. He, C. Xu, and L. Tan, "Sequential human activity recognition based on deep convolutional network

and extreme learning machine using wearable sensors," *Journal of Sensors*, vol. 2018, 2018.

[24] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using binary motion image and deep learning," *Procedia computer science*, vol. 58, pp. 178–185, 2015.

[25] O. Nafea, W. Abdul, G. Muhammad, and M. Alsulaiman, "Sensor-based human activity recognition with spatio-temporal deep learning," *Sensors*, vol. 21, no. 6, p. 2141, 2021.

[26] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input cnn-gru based human activity recognition using wearable sensors," *Computing*, vol. 103, no. 7, pp. 1461–1478, 2021.

[27] P. U. Putra, K. Shima, and K. Shimatani, "A deep neural network model for multi-view human activity recognition," *PloS one*, vol. 17, no. 1, p. e0262181, 2022.

[28] D. Thakur, S. Biswas, E. S. Ho, and S. Chattopadhyay, "Convae-lstm: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition," *IEEE Access*, 2022.

[29] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "Human activity classification using the 3dcnn architecture," *Applied Sciences*, vol. 12, no. 2, p. 931, 2022.

[30] D. Anguita, A. Ghio, L. Oneto, X. Parra Perez, and J. L. Reyes Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, 2013, pp. 437–442.

[31] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 2010, pp. 233–240.

[32] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, 2012, pp. 1–8.

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

[34] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 42–50, 2008.

[35] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–33, 2014.

[36] M. Zhang and A. A. Sawchuk, "Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 1036–1043.

[37] D. Anguita, A. Ghio, L. Oneto, X. Parra Perez, and J. L. Reyes Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, 2013, pp. 437–442.

[38] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in *Proceedings of the 13th ACM conference on embedded networked sensor systems*, 2015, pp. 127–140.

[39] developers.google, "Classification: Accuracy nbsp;—nbsp; machine learning crash course nbsp;—nbsp; google developers." [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy

[40] A. G. D'Sa and B. Prasad, "A survey on vision based activity recognition, its applications and challenges," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019, pp. 1–8.

[41] A. Drosou, D. Ioannidis, K. Moustakas, and D. Tzovaras, "Spatiotemporal analysis of human activities for biometric authentication," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 411–421, 2012.

[42] R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," *Artificial Intelligence Review*, vol. 50, no. 2, pp. 283–339, 2018.

[43] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: a review," *Advances in computational intelligence systems*, pp. 341–371, 2017.

[44] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[45] M. Youssef and V. Asari, "Human action recognition using hull convexity defect features with multi-modality setups," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1971–1979, 2013.

[46] A. D. Antar, M. Ahmed, and M. A. R. Ahad, "Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2019, pp. 134–139.

Aniruddh G. Fataniya Aniruddh G. Fataniya is working as an Assistant Professor at Department of Computer Engineering of Charotar University of Science and Technology, Changa, Gujarat, India. He has received the B.E. degree in Computer Engineering from Gujarat University and M.E. degree in Computer Science and Engineering from Gujarat Technological University. He is pursuing Ph.D. from Charotar University of Science and Technology. He has guided more than 5 Post Graduate students and 20 Under Graduate students for their project and research work. His research interests are Computer Vision, Deep Learning and new futuristic technologies.

Dr. Hardik P. Modi Dr. Hardik P. Modi is working as Associate Professor at Department of Electronics and Communication Engineering of Charotar University of Science Technology, Changa, Gujarat, India. He has received the B.E. degree in Electronics and Communication Engineering from Sardar Patel University, M.E. in Electronics and Communication Systems Engineering from Dharmsinh Desai University and Ph.D. degree from Charotar University of Science and Technology. He has filed one patent and published more than 50 research papers in peer reviewed journals. He has guided more than 20 Post Graduate students and 50 Under Graduate students for their project and research work. His research interests are bio medical image processing, computer vision, wireless communication and new futuristic technologies.