Hybrid LRCN Approach for Human Activity Recognition from Video Data

Abstract: In today's world, our everyday life is significantly facilitated by the extensive use of technology. Particularly, the use of artificial intelligence has enhanced the ability to assist in various daily activities. One notable application of AI is the Human Activity Recognition (HAR) system, which employs Artificial Neural Networks (ANN) to understand and classify human actions based on real time training data. In this study, we present an LRCN (Long-term Recurrent Convolutional Network) model-based Activity Recognition system that is designed to detect human activities. The model utilizes a vast collection of videos from the UCF-50 dataset to create a comprehensive statistical model. It learns spatiotemporal features that makes it robust than other deep learning models. We compared the LRCN approach with two well-known models, CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) proposed in the past researches. From the comparison of these three different models, the LRCN model demonstrated higher accuracy in activity detection.

Keyword: HAR, UCF-50, LRCN, Accuracy.

I. INTRODUCTION

Human Activity Recognition (HAR) using mobile sensors and Long-term Recurrent Convolutional Network (LRCN) model represents a sophisticated approach to discerning and interpreting human actions from smartphone sensor data. With the proliferation of smartphones equipped with a variety of sensors like accelerometers and gyroscopes, the opportunity to infer users' activities in real-time has become increasingly feasible.

[5] Motion can be measured using an accelerometer, a gyroscope, or a magnetometer. These are packed similarly to other integrated circuits and can give either digital or analog outputs. Embedded within smartphones, these can continuously capture the dynamics of user's movements and orientations, generating rich streams of sensor data. This data, when properly processed and analyzed, can reveal insights into various activities such as walking, running, cycling, and more, making smartphones a ubiquitous platform for HAR applications.

[7] The integration of LRCN with mobile sensors presents a compelling approach for capturing the spatial and temporal patterns inherent in sensor data streams. LRCNs, which combine convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs) for temporal sequence modeling, for handling the dynamic and multidimensional nature of mobile sensor data.

[8] By leveraging the strengths of both CNNs in spatial feature extraction and LSTMs in sequential modeling, LRCN models can effectively learn complex patterns and variations in sensor data sequences. This allows for accurate recognition and classification of diverse human activities, even in the presence of noise and variability in sensor readings.

[10] The application of LRCN models in HAR holds immense promise across various domains, including healthcare, fitness tracking, augmented reality (AR), and context-aware computing. These models can facilitate personalized activity monitoring interfaces, enhancing user experiences and well-being.

In this research, we present a comprehensive exploration of HAR using CNN, LSTM and LRCN models with mobile sensors. We discuss the methodology for data collection and preprocessing, the architecture of these models, experimental results demonstrating its efficacy, and potential implications for real-world applications.

II. LITERATURE REVIEW

In recent times, Human Activity recognition system is being researched and implemented more. It's because of the advancement in technology specifically in mobile sensors which uses accelerometer to sense body movements. Previously, accelerometers were used on subject's body for research process since mobile sensors were less prevalent even a decades ago.

Prasad, Ashwani, et al. [4] designed a 2D CNN model for recognizing and predicting human activities based on accelerometer readings. The results were compared with renowned LSTM model. The study showed the importance of learning before model development as the learning scores can be compared with predicted scores.

Mahendra R [6] developed something that overcome the disadvantages of mobile sensors. Error was reduced caused by the usage of sensors by eliminating them. Logistic regression/Logistic classification model was used to detect human activity actions. LRCN and ConvLSTM, these two models were implemented. Since the accuracy can never be 100%, so the results will always have a possibility of incorrect output.

In [9], Sajib Uzzaman developed a ConvLSTM and LRCN based model to detect human activities. Two datasets UCF50 and HMDB51 were used in the paper. Though the CNN model showed higher accuracy in both datasets when used as training model, LRCN

model constructed a real time Human Activity Recognition System and had better results than ConvLSTM model.

Shrinathika, Chamani [13] proposed an approach to predict human activities by the developed CNN and LSTM model. WISDOM dataset was used in the paper. Using perfect model hyper parameters in the networks of the two models made them robust and fast. Moreover, the author proposed the activity recognition system as a solution for childcare and eldercare monitoring system based on IoT. It can also be used in other datasets generated by sensors.

In our proposed system we used Long-term Intermittent Convolutional Network (LRCN), a single model that mixes LSTM and CNN layers. For temporal sequence modelling, the extracted spatial features are given to the LSTM subcaste (s) at each time step, while the convolutional layers are utilized for spatial point birth from the frames. As we claw into the specifics of the design, the integration of LRCN styles with our dataset holds great eventuality for achieving high-performance mortal exertion recognition on smartphones.

III. METHODOLOGY

In our approach, the implementations were carried out under four main areas. First, we developed basic LRCN model and also CNN, LSTM model separately for the activity recognition. Next, we trained the models for our dataset. At the third step, we developed the CNN, LSTM and LRCN model for the classification and the prediction of the activities. Finally, we plotted accuracy and loss of data and also the Confusion Matrix. The LRCN model combined CNN and LSTM layers together. This way the network learns spatiotemporal features directly in an end-to-end training, resulting in a robust model. The workflow of the LRCN model is shown for better understanding:

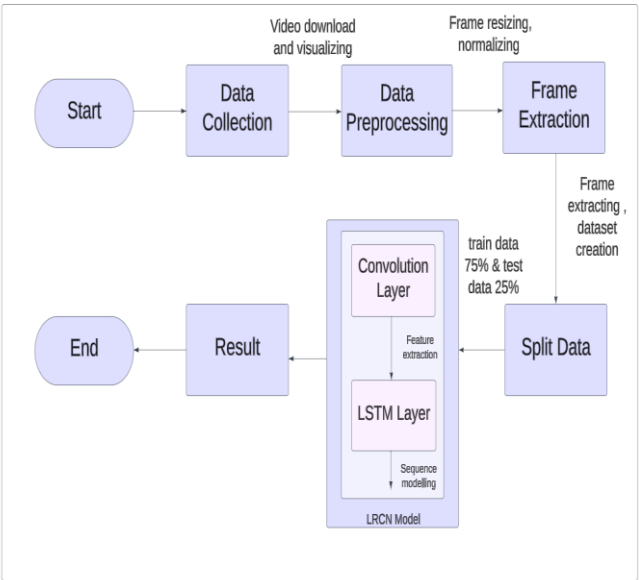

Figure 1: Flowchart

1.Input Dataset: We've used UCF50- Action recognition dataset to train our models. The conditioning in the datasets include swing, running, play violin, biking, diving, pullups playing tennis etc. similar data can be captured by an on- board computer system, or a variety of surveillance cameras. The dataset contains:

- 50 Action Categories
- 25 Groups of videos per action category
- 133 average videos per category

Figure 2: Examples of dataset clips

2.Preprocessing of Dataset: To prepare the dataset, we undertook some preprocessing. Before normalizing the data to the range [0-1] by splitting the value of every pixel by 255, we read the video clips from the dataset, shrunk the video frames to a fixed length and height of 64, and reduced the computations. This allows the network to train more quickly.

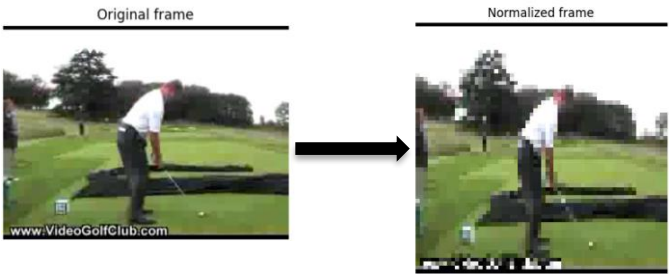

Figure 3: Resized and Normalized frame

3.Split Dataset into Train and Test: The created dataset was split into training (75%) and testing (25%) sets. Before splitting, we shuffled the dataset to remove bias and produce splits that accurately reflected a distribution of the data as a whole.

4.Model Construction: We used three deep learning models:

- **Convolutional Neural Network Approach:** CNNs are primarily used for image recognition

and bracket tasks. They consist of convolutional layers that apply pollutants to input images to prize features CNNs are extensively used in computer vision tasks like object discovery, facial recognition, and image segmentation.

- **Long-Short Term Memory Approach:** LSTMs are a type of intermittent neural network designed to handle successional data like textbook, speech, and timeseries. They address the evaporating grade problem of traditional RNNs by introducing gating mechanisms.

- **Long-Term Recurrent Convolutional (LRCN) Approach:** We applied the Convolution model to every frame in our LRCN architecture using time-distributed Conv2D layers. An LSTM layer will get the feature that was taken from the Conv2D layers after it has been smoothed using the Flatten layer. After that, the LSTM layer's output will be used by the Dense layer.

5.Model Visualization: The LRCN model architecture is constructed and shown below:

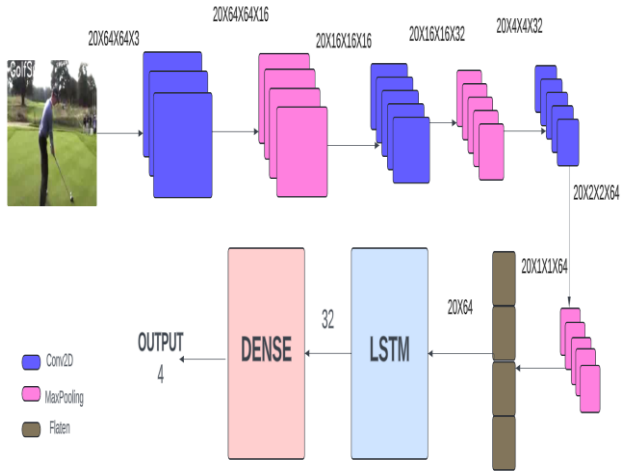

Figure 4: LRCN Model Architecture

The LRCN model architecture processes input video frames of shape 20x64x64x3 through several Conv3D layers, which sequentially transform the frames to 20x16x16x16, 20x4x4x32, and finally to 20x1x1x64. These features passed through a flatten layer, reducing the dimensionality to 20x64 to make the data more manageable and to highlight essential features. The vectors are then fed into an LSTM layer, which captures temporal dependencies to 32 but memory. The dense layer has 4 neurons as we trained the model for 4 chosen classes. The model concludes with an output layer that classifies the video into one of four possible categories.

IV. Exploratory Setup and Evaluation

1.Experiment: The experiment is conducted using the UCF-50 dataset. The data set was first split into two orders training and testing, with 75% and 25% of the videos in each. Prior to the split, in order to minimize bias and generate splits that accurately depict the distribution of the data overall, the dataset is further shuffled.
Videos were converted into single frames to train and analyze performance of the models. We have compared CNN, LSTM and LRCN model performance according to accuracy.

2.1. Result of CNN Approach: Figure 8 shows the training and validation accuracy as well as the training and validation loss for a CNN model. Table 1 shows the level of accuracy of the model. The accuracy of this model is greater than LSTM but less than LRCN:

Pooling	Activation Function	Layer	Dropout	Result
Maxpooling2D+ GlobalAveragePooling1D	Relu+Softmax	2	25%	87.7%

Table1: Accuracy Result of CNN Approach

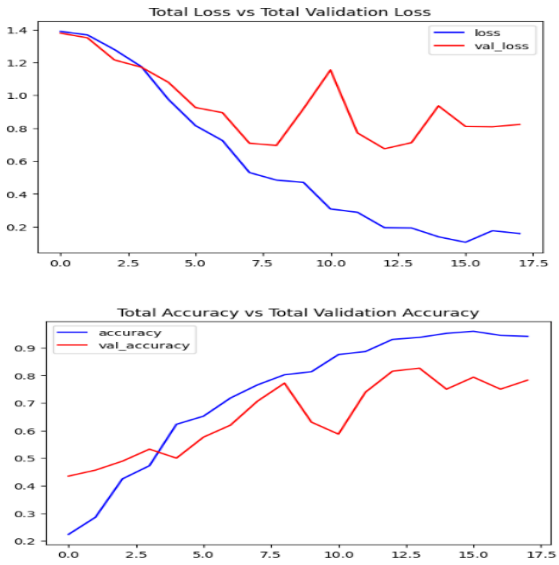

Figure 5: CNN Model Accuracy & Loss

CNN training progressed at satisfactory level. While learning improved (lower training loss, higher training accuracy), a validation loss uptick at the end suggests potential overfitting.

2.2. Result of LSTM Approach: Figure 9 shows the training and validation accuracy as well as the training and validation loss for a LSTM model. Table 2 shows the level of accuracy of the model. Here the accuracy of LSTM is less than both CNN and LRCN:

Pooling	Activation Function	Dropout	Result
Maxpooling3D	tanh	20%	45.65%
Maxpooling3D	tanh	20%	58.67%
Maxpooling3D	tanh	20%	73.91%
Average Pooling	Softmax	20%	76.23%

Table2: Accuracy Result of LSTM Approach

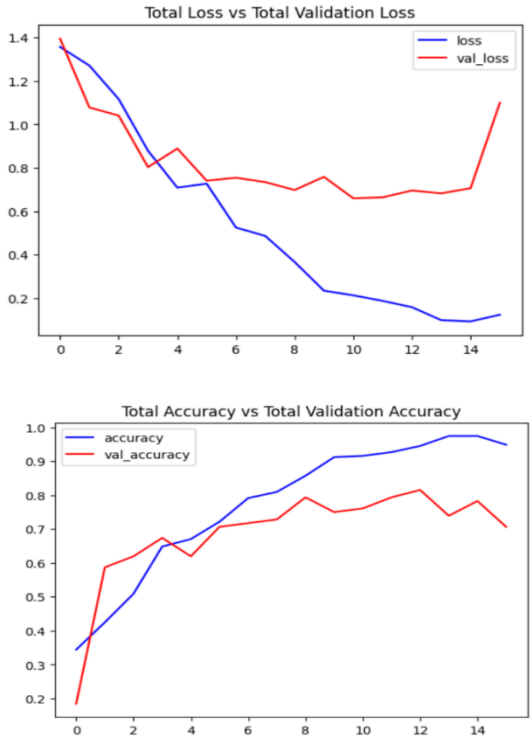

Figure 6: LSTM Model Accuracy & Loss

The LSTM model's training loss consistently decreases, but validation loss increases after epoch 10, indicating overfitting. Training accuracy improves steadily, while validation accuracy fluctuates, showing less stable generalization performance.

2.3. Result of LRCN Approach: Figure 10 shows the training and validation accuracy as well as the training and validation loss for a LRCN model. Table 3 shows the level of accuracy of the model. Here the accuracy of LRCN is more than both CNN and LRCN which means it's the most accurate out of three:

Pooling	Activation Function	Dropout	Result
Maxpooling2D	Relu	25%	88.04%
Maxpooling2D	Relu	25%	89.13%
Maxpooling2D	Relu	25%	85.87%
Average Pooling	Softmax	25%	91.80%

Table3: Accuracy Result of LRCN Approach

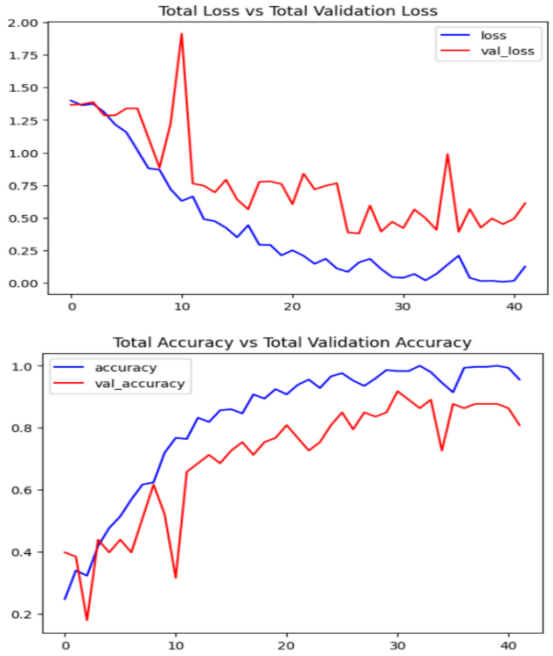

Figure 7: LRCN Model Accuracy & Loss

Training in LRCN exhibits its ability. Lower training loss and higher training accuracy indicate improved learning; nevertheless, a final validation loss bump raises the possibility of overfitting.

4.Result Comparison:

Model	**Accuracy**
CNN	86.89%
LSTM	76.23%
LRCN	91.80%

Table 4: Accuracy of the models

5.Confusion Matrix and Classification Report: The confusion matrix and Classification Report of all the models where 4 classes "WalkingWithDog", "TaiChi", "Swing", "HorseRace" labeled as 0,1,2,3 used, is represented below:

CNN:

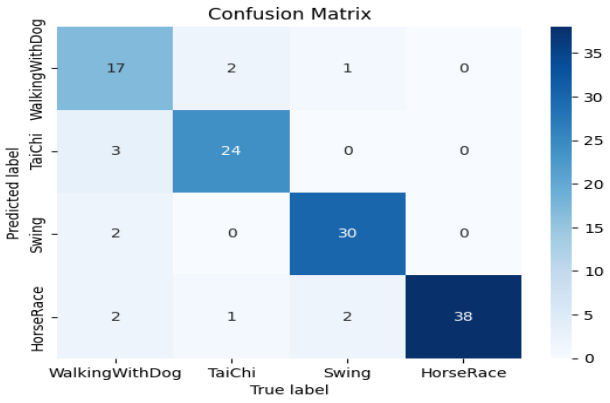

Figure 8: CNN Model Confusion Matrix

The confusion matrix, where the y-axis represents predicted labels and the x-axis represents true labels, displays how well the CNN model performed in predicting tasks. Off-diagonal cells display misclassifications, whereas diagonal cells show accurate predictions. Here, the model correctly predicted 17 instances of label 0 but misclassified 3 as 1, 2 as label 2 and 2 as label 3. Similarly, for label 1, it made 24 correct predictions but misclassified 2 as label 0 and 1 as label 3. Label 2 had 30 correct predictions but notable misclassifications as label 0 (1 instances) and label 3 (2 instances). The model performed best with label 3, correctly predicting 38 instances with no misclassification. Overall, the model shows high accuracy but has room for improvement.

Class	Precision	Recall	F1 Score	Support
0	0.85	0.75	0.77	24
1	0.89	0.89	0.89	27
2	0.94	0.91	0.92	3
3	0.88	1.00	0.94	38

Table 5: CNN Model Classification Report

The classification report shows precision, recall, f1 score and support of each class for the CNN approach. The precision was highest for class 2 and recall, f1 score was best for class 3. The overall precision and recall were good for all the classes.

LSTM:

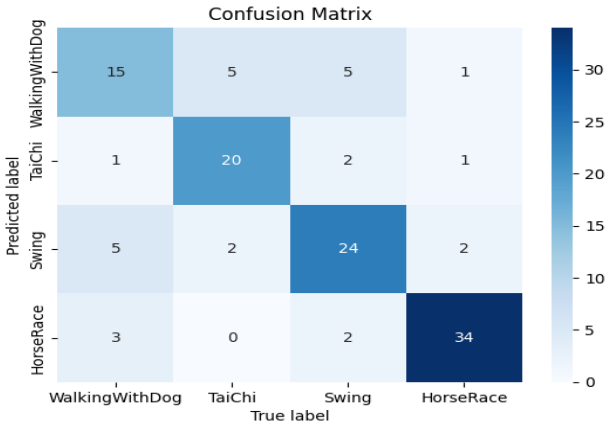

Figure 9: LSTM Model Confusion Matrix

The confusion matrix illustrates the performance of an LSTM model in predicting human activities, with true labels on the x-axis and predicted labels on the y-axis. The diagonal cells show accurate predictions., showing that the model correctly identified 15 instances of label 0, 24 of label 1, 31 of label 2, and 28 of label 3. Misclassifications are shown in the off-diagonal cells, where label 0 was often misclassified as label 2 (7 instances), and label 3 had multiple misclassifications, particularly 5 instances as label 0 and 4 instances as label 2. While the model shows good accuracy for labels 1 and 2, it struggles more with labels 0 and 3, indicating areas for improvement to reduce these misclassifications.

Class	Precision	Recall	F1 Score	Support
0	0.58	0.62	0.60	24
1	0.83	0.74	0.78	27
2	0.73	0.73	0.73	3
3	0.87	0.89	0.88	38

Table 6: LSTM Model Classification Report

The classification report shows precision, recall, f1 score and support of each class for the LSTM approach. The precision, recall and f1 score highest for class 3 lowest for class 0. The overall precision and recall were not satisfactory for all the classes except class 3.

LRCN:

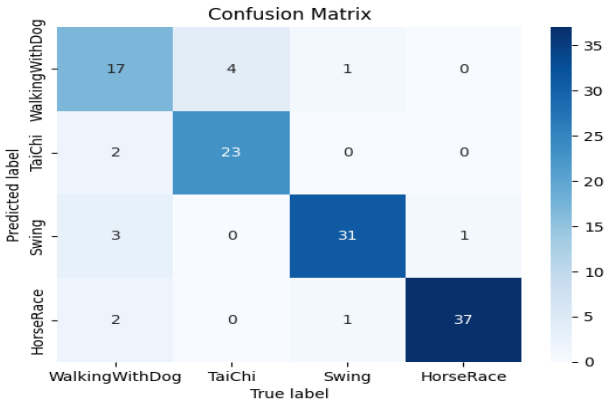

Figure 10: LRCN Model Confusion Matrix

The confusion matrix for the LRCN model illustrates its performance in predicting human activities, with true labels on the x-axis and predicted labels on the y-axis. Correct predictions are shown along the diagonal, where the model accurately identified 17 instances of label 0, 23 of label 1, 31 of label 2, and 37 of label 3. Misclassifications include 3 instances of label 0 being predicted as label 2, 4 instances of label 1 predicted as label 0, and 1 instances of label 3 predicted as label 2. Overall, the LRCN model shows strong performance with high accuracy for labels 2 and 3, there is room for improvement in reducing the misclassifications of label 0 and 1.

Class	Precision	Recall	F1 Score	Support
0	0.77	0.71	0.74	24
1	0.92	0.85	0.88	27
2	0.89	0.94	0.91	3
3	0.93	0.97	0.95	38

Table 7: LRCN Model Classification Report

The classification report shows precision, recall, f1 score and support of all classes for the LSTM approach. The precision, recall and f1 score were best for class 3 worst for class 0. The overall precision and recall were very good except class 0.

V. CONCLUSION

This is a thorough investigation of Human Activity Recognition (HAR) systems with an emphasis on accurately detecting activity through the use of a Long-term Recurrent Convolutional Network (LRCN) model. Through experimentation with the UCF50 dataset, the LRCN model demonstrates superior performance compared to standalone Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, achieving the highest accuracy of 91.80%. This highlights how deep learning approaches, in particular the LRCN architecture, might advance HAR systems for a range of uses. The LRCN model can efficiently extract spatial features from sensor data and model temporal sequences thanks to the merging of CNN and LSTM. This opens up new possibilities for context-aware computing, fitness tracking, and improved healthcare monitoring. Further research could focus on optimizing the LRCN model and exploring its applicability in real-world scenarios to enhance user experiences and well-being.

VI. REFERENCE

[1] Benzyane M, Azrour M, Zeroual I, Agoujil S. "Investigating the Influence of Convolutional Operations on LSTM Networks in Video Classification." Data and Metadata. 2023:2:152.

[2] Vrigkas,Michalis. "A Review of Human Activity Recognition Methods." Frontiers in Robotics and AI. Vol 2. 2015.

[3] Pandya, Meet, Abhishek Pillai, and Himanshu Rupani. "Segregating and Recognizing Human Actions from Video Footages Using LRCN Technique." Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020. Springer Singapore, 2021.

[4] Prasad, Ashwani, et al. "Human activity recognition using cell phone-based accelerometer and convolutional neural network." Applied Sciences 11.24 (2021): 12099.

[5] Aboo, Adeeba Kh, and Laheeb M. Ibrahim. "Human Activity Recognition Using a Hybrid CNN-LSTM Deep Neural Network." Webology (ISSN: 1735-188X) 19.1 (2022).

[6] Mahendra R, Vidyarani, H. J., and Chandrakanth G. Pujari. "HUMAN ACTIVITY RECOGNITION."(Eissn: 2582-5208), IRJETS, 2022.

[7] Domingo, Jaime Duque, Jaime Gómez-García-Bermejo, and Eduardo Zalama. "Improving human activity recognition integrating lstm with different data sources: Features, object detection and skeleton tracking." IEEE Access 10 (2022): 68213-68230.

[8] Keshinro, Babatunde, Younho Seong, and Sun Yi. "Deep Learning-based human activity recognition using RGB images in Human-robot collaboration." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 66. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2022.

[9] Uzzaman MS, Debnath C, Uddin MA, Islam MM, Talukder MA, Parvez S. LRCN based human activity recognition from video data. SSRN Electronic Journal. 2022.

[10] Yang, Jianbo, et al. "Deep convolutional neural networks on multichannel time series for human activity recognition." Ijcai. Vol. 15. 2015.

[11] Zeng, Ming, et al. "Convolutional neural networks for human activity recognition using mobile sensors." 6th international conference on mobile computing, applications and services. IEEE, 2014.

[12] Xu, Cheng, et al. "InnoHAR: A deep neural network for complex human activity recognition." Ieee Access 7 (2019): 9893-9902.

[13] Shrinathika, Chamani, Huei-Ling Chiu. Human Activity Recognition using CNN & LSTM. IEEE 20XX.

[14] Bevilacqua, Antonio, et al. "Human activity recognition with convolutional neural networks." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III 18. Springer International Publishing, 2019.

[15] Ha, S., Choi, S.: "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors." In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 381–388, July 2016.

[16] Alsheikh, M.A., Selim, A., Niyato, D., Doyle, L., Lin, S., Tan, H.P.: "Deep activity recognition models with triaxial accelerometers." CoRR abs/1511.04664 (2015).

[17] Sansano, Emilio, Raúl Montoliu, and Oscar Belmonte Fernandez. "A study of deep neural networks for human activity recognition." Computational Intelligence 36.3 (2020): 1113-1139.

[18] Lara, Oscar D., and Miguel A. Labrador. "A survey on human activity recognition using wearable sensors." IEEE communications surveys & tutorials 15.3 (2012): 1192-1209.

[19] Ordóñez, Francisco Javier, and Daniel Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition." Sensors 16.1 (2016): 115.

[20] Wei, Li, and Shishir K. Shah. "Human Activity Recognition using Deep Neural Network with Contextual Information." VISIGRAPP (5: VISAPP). 2017.