Vision-based Human Activity Recognition Uses a Deep Learning Approach

Pranta Kumar Sarkar
Computer Science and Engineering,
Bangabandhu Sheikh Mujibur Rahman Science and Technology University,
Gopalganj, Bangladesh

Moskura Hoque
Computer Science and Engineering,
Mawlana Bhashani Science and Technology University Santosh,
Tangail, Bangladesh.

Mostofa Kamal Nasir, PhD
Computer Science and Engineering,
Mawlana Bhashani Science and Technology University Tangail,
Bangladesh

ABSTRACT
In today's world, daily life increasingly depends on vision-based advanced technologies, which enhance the reliability and convenience of human lifestyles. Among these technologies, vision-based Human Activity Recognition (HAR) stands out as a comprehensive and challenging field of study, with broad exploration and practical applications. HAR systems are designed to identify diverse human actions under varying environmental conditions.Vision-based activity recognition plays a crucial role in a wide range of applications, including user interface design, robot learning, security surveillance, healthcare, video searching, abnormal activity detection, and human-computer interaction. This study focuses on recognizing various human activities in real-world settings, highlighting the importance of consistency and credibility in the results.To achieve this, data was collected from multiple sources and processed using three distinct models—Convolutional Neural Network (CNN), VGG-16, and ResNet50—to identify the most effective approach for activity recognition. Among these, a specific architectural CNN model was further evaluated for its ability to capture human activity features in specific video sequences. The training, validation, and testing phases utilized a comprehensive dataset comprising 56,690 images. Remarkably, the proposed system achieved an impressive accuracy of 96.23% after 30 epoch running and low validation loss illustrate its effectively recognition each feature.

Keywords
Computer Vision, Activity Recognition, CNN, Deep Learning, High Performance.

1. INTRODUCTION
Vision-based human activity recognition (HAR) is a significant computational system in the field of computer vision analysis, drawing substantial attention from both academic and industrial sectors. Its primary goal is to achieve consistent labeling of identical activities, even when performed by different individuals in various settings or contexts. Vision-based HAR [1] systems make an effort to automatically identify and analyze such HAs using the data collected from various kinds of sensors. The main applications include privacy, visual monitoring, video acquisition, entertainment, and anomalous activity monitoring to catch glimpses of what's happening and to identify potentially harmful or unlawful activities. HAR also aids in enhancing Human-Computer Interaction capabilities, such as the automatic situation recognition. Utilizing deep learning techniques enables the achievement of tasks related to human activity identification, harnessing their powerful capability to extract significant features from raw data. However, use the deep learning method for HAR required to design and select relevant features [2]. Since it is impractical to gather a significant amount of labeled activity data, it has the capacity to learn from unlabeled data, which is significant and helpful for HAR [3,4]. Researchers have been studying activity categorization and pattern recognition methods for a long time [5,6], with an emphasis on the design and a few specific elements that are task-dependent. Recently, there has been extensive interest in deep learning and machine learning approaches [7-9] for a diverse range of problems including model building, dataset development, covering all situations, and system analysis less efficiently. Automatically, deep learning extracts similar patterns or attributes from CCTV footage [10]. This section discusses related studies conducted in related domains and how they influenced the fundamental organization of this work. Following is a list of the several pertinent variables that were considered in the results from the various 2052 articles [11,12] processed that most of Science Direct and IEEE Xplorer where more than 256 articles publish [1,8,13], far surpassing other specialized databases such as Scopus, Web of Science, and ACM: (1) the year of publication of the article see Fig. 1. shows that of the total articles, 64% of the studied ones mention conference publications, 4 mention books, and 36% mention journals. If they can be identified using HAR system, the majority of human everyday jobs can be mechanized or made simpler. HAR systems are frequently either supervised or unsupervised [14]. To analyze surveillance footage, a variety of modules are employed, such as object detection, action recognition, pattern recognition, and categorization of observed activities into categories like anomalous, abnormal, or normal [15,16]. HAs that utilize the data gathered from the various types of sensors [17]. It participates in the creation of a variety of significant applications, including those related to human computer interaction (HCI) [20,21], virtual reality, security [15-18], video surveillance, and home monitoring [19-21]. In the authors system use trained frames from the videos will be the output that the neural network provides after training 56690 images make up the dataset used in this article. Results indicate that the performance of the recognition system has improved in comparison to the preceding system.

The structure of the paper is as follows: Section 1 provides a brief introduction to human activity recognition, its applications in various contexts, and related work. Section 2 outlines the proposed model, the dataset, and their functionality. Section 3 presents a discussion of the results. Finally, Section 4 concludes the paper and highlights some limitations of the study.

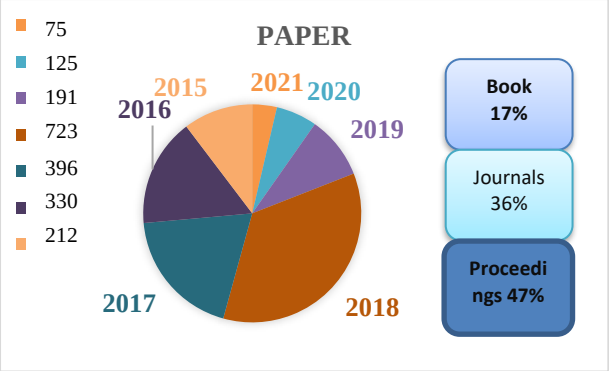

Fig. 1: Survey on HAR advancements

2. PROPOSED METHODOLOGY

This paper proposed a system, The authors produced a fresh dataset through a procedure, and this work used a deep learning-based model to track and identify people moving about outside. The internal organization of the proposed work, proposed dataset, and comparative study are briefly described in this section. HAR approaches according to the recognition process are divided into three important stages: (1) Data collection and preprocessing (2) Model building and training (3) Activity classification. The overall workflow for deep learning-based human activity classification is illustrated Fig 2.

The methodology this experiment utilize is broken down into the following steps.

• Surveillance regions are monitored, and videos are collected as input.

• From these input videos, frames capturing multiple actions are extracted as images from various sources.

• Further data preprocessing is performed on the extracted frames to improve performance.

• Multiple human activity datasets are used for training and testing with various deep learning models.

• The model that best fits the dataset is selected, yielding results that demonstrate an improved recognition rate compared to previous methods.

• Finally, the chosen model is employed to detect and classify human activities.

2.1 Feature Extraction Model

In Therefore, datasets with images are also beneficial. The datasets are frequently utilized to construct numerous types of applications. Artificial neural networks called convolutional neural networks (CNN) were created with the purpose of processing structured matrices, such as images[22]. Additionally, choose appropriate some parameter such as learning rate 0.0001, batch size 64, 35 of epochs, and ReLU and SoftMax activation functions significantly influences training outcomes. Optimization strategies, including the use of algorithms like Adam further enhance the model's ability to converge effectively. Together, these design elements and tuning strategies determine the model's accuracy, generalization ability, and robustness in real-world applications. in Fig.3.

These approaches are inappropriate for low-power devices because of their computational complexity needs as well as learning rate slower training but more accurate work .

Following are the steps used in this system to carry out the process:

- Pre-processing
- Feature extraction
- Object tracking
- Behavior Understanding

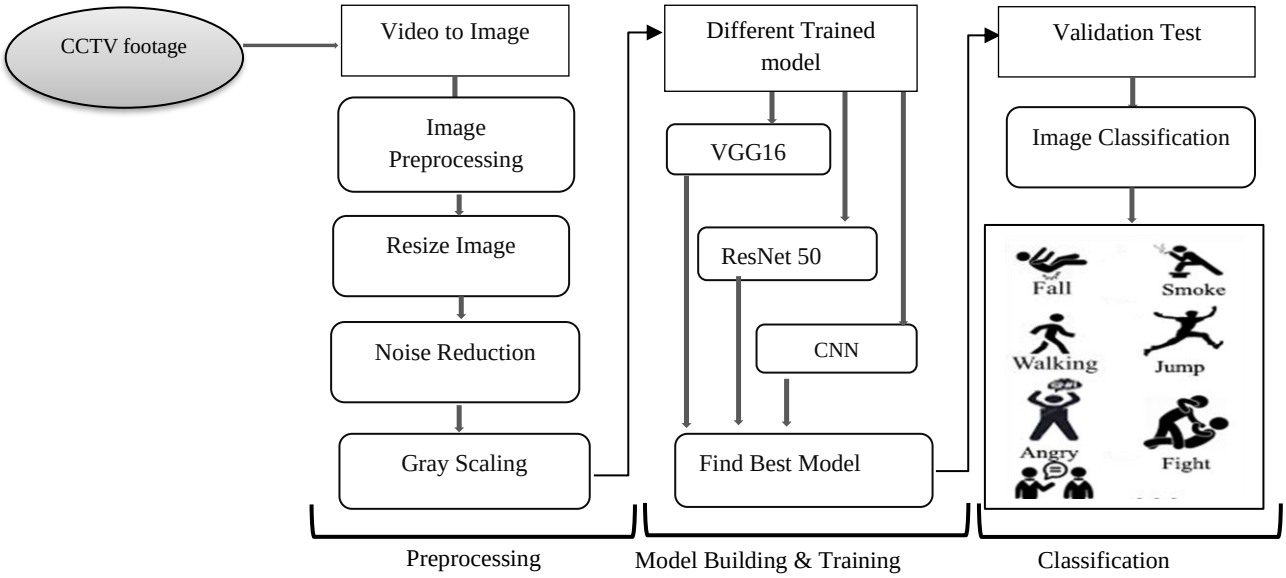

Fig. 2: Design Layout of the System

This model architecture is composed of four Conv3D layers [23], two max-pooling layers, and two completely connected layers. With the default settings, this work apply 32 filters to the first pair of Conv3D layers and 64 filters to the second pair. All kernels in the model are 3 x 3 x 3, and the loss computation is done using the Adam optimizer and cross-entropy. At the model's finish, it has two dense layers with 128 and 11 neurons, respectively. The output, which is multiclass, represents 11 separate human activities.

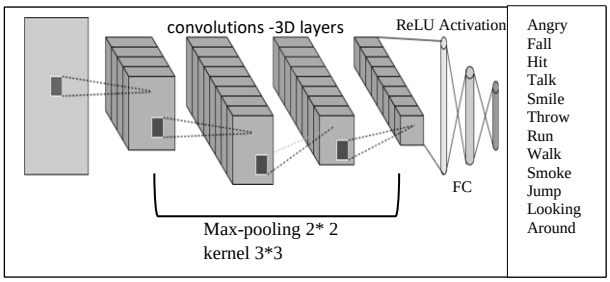

Fig. 3: Neural Network Architecture

2.2 Dataset:

In many available sources, human activity recognition data often relies on single-frame object analysis to detect activities. However, this approach can lead to reduced accuracy, as it fails to capture the continuous flow of movement over time. Moreover, in certain local areas, the improper installation of surveillance cameras poses additional challenges for accurate object detection. In such cases, datasets play a vital role in supporting research by helping evaluate the effectiveness and accuracy of both existing and emerging methods.

After obtaining a substantial number of photographs, this work conducted a manual curation process to eliminate irrelevant images. This procedure was applied uniformly to all 11 classes. As presented in Table 1, The experiment's dataset comprises a total of 56,690 photos.

Table 1. Summary of Dataset Features

Actions	11
Video Clips	300
Each action Videos	12-17
Total Duration	2500 sec
Frame Rate	30fps
Video Resolution	480 x 720
Total image	56690

Initially, data was collected from diverse websites and social media platforms based on predetermined criteria also use noise reduction Gaussian Blur technique. The image is first resized to dimensions of 224 × 224 pixels and converted to a single-channel grayscale format. Next, the grayscale image is transformed into a binary image using a threshold value of 127. Finally, edge detection is performed using the Canny method to emphasize key features and object boundaries, as shown in Fig 4. This step significantly enhances the accuracy of activity recognition.

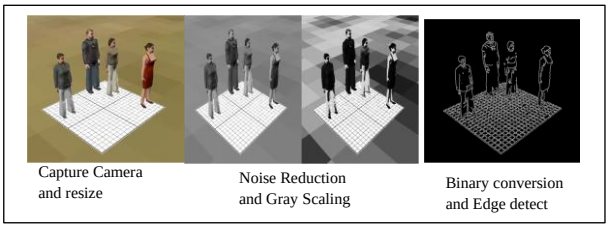

Fig. 4: Steps Involved in Data Preprocessing

These datasets include specialized features for a variety of research needs, including gestures, suspicious human behavior, abnormality, visual surveillance, video acquisition, amusement, and monitoring of aberrant activity.

3. EXPERIMENTAL RESULTS

To compare the suggested human activity recognition technique's effectiveness with other methods and the performance of the method on complicated daily activity data. This section includes a number of experiments to assess how well deep learning model classifiers perform on various data sets. This work's design assesses the efficacy and efficiency of Activity recognition performance. A deep learning strategy was utilized in this experiment to complete this portion, which seeks to identify and categorize human activity in certain daily human behavior. Modern algorithms still struggle with a number of issues that lead to incorrectly categorizing behaviors or activities, despite great improvements in computer vision HAR machine accomplishments.

3.1 Test Data Performance Of Activity Classification Use Confusion Matrix

The appropriate weights were achieved after training using 75% of the dataset. The remaining 20% of the dataset was evaluated using these weights, providing a thorough evaluation of the models' accuracy and precision that was recorded for future comparison and ultimately used to identify the top detection model. Fig. 5. displays the confusion matrix of the CNN classifier on the study's development dataset.

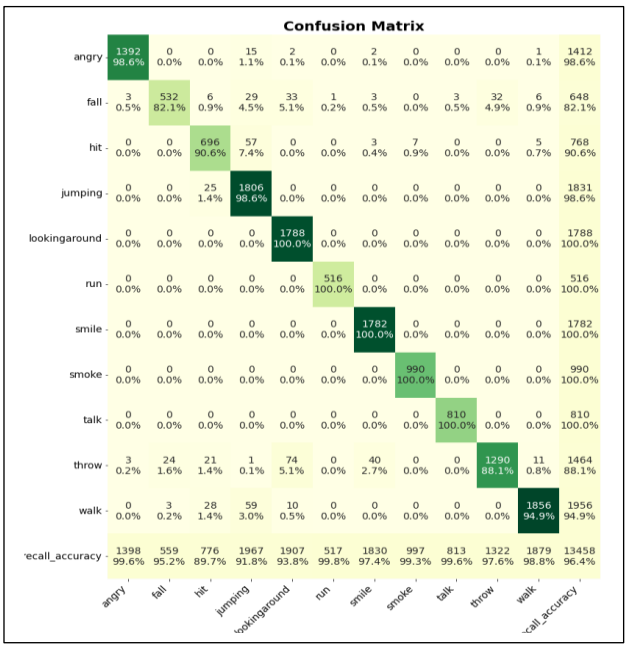

Fig. 5. Assessment of Classification Results via Confusion Matrix

Correct predictions are displayed across the diagonal cells of the table, and the majority of action classes, such as anger, fall, hit, jumping, looking around, running, smiling, smoking, talking, throwing, and walking, are nearly predicted well with greater than 90% accuracy. The study have chosen a more realistic feature extraction effect on this work.

3.2 Metrics for Evaluation in an Object Detection Model

Mean Average Precision (mAP), which is utilized in computer vision, is used to assess the Object Detection Model [24]. The number of accurate predictions the model made is used to determine accuracy.

• Accuracy: The ratio of the number of accurate forecasts to the total number of TP estimates the model produced.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, FN, and FP, respectively, stand for "True Positive Class Prediction," "True Negative Class Prediction," and "True Negative Class Prediction."

• Precision is the ratio of real true forecasts to all of the model's actual true predictions.

$$\text{Precision} = \frac{1}{c}\left(\sum_{c=1}^{c} \frac{\text{tp}_c}{\text{tp}_c + \text{fp}_c}\right)$$

where C stands for the total number of classes, tpc for true positives for a specific class, and fpc for false positives for a specific class.

• Recall: the proportion of accurate predictions made by the model to the actual number of accurate forecasts.

$$\text{Recall} = \frac{1}{c}\left(\sum_{c=1}^{c} \frac{\text{tp}_c}{\text{tp}_c + \text{fn}_c}\right)$$

• Harmonic mean of the estimated accuracy and recall is the F1-score.

$$\text{F1} - \text{Score} = \sum_{c=1}^{c} \frac{\text{precision}_c * \text{recall}_c}{\text{precision}_c + \text{recall}_c} * 2\left(\frac{\text{n}_c}{\text{N}}\right)$$

where N is the overall sample count, nc is the number of samples in class c, precision is the precision value for that class c, and recall is the recall value for that class [25].

In this study, Evaluated the performance of each model using three different evaluation matrices, including Accuracy, Precision, and Recall. Accuracy, precision, recall, and f1-score, four commonly used assessment measures, are used to gauge the model's effectiveness those all value show table 2. The total number of images obtained from depth data were

Table 2. Deep Learning Model Precision, Recall, F-1 Score

Model	Precision	Recall	F-1 Score
VGG-16	0.93	0.91	0.91
ResNet50	0.86	0.83	0.83
CNN	0.97	0.96	0.96

56690, of which 37,000 samples were selected for training, 6500 for validation and 13965 for testing. ResNet-15, VGG-16, and CNN approaches were all tested. The training and validation accuracy of at least 30 epoch execute results are shown in Fig. 6. With an accuracy rate of 96.4%, CNN achieves the top results in this study's traditional deep learning category.

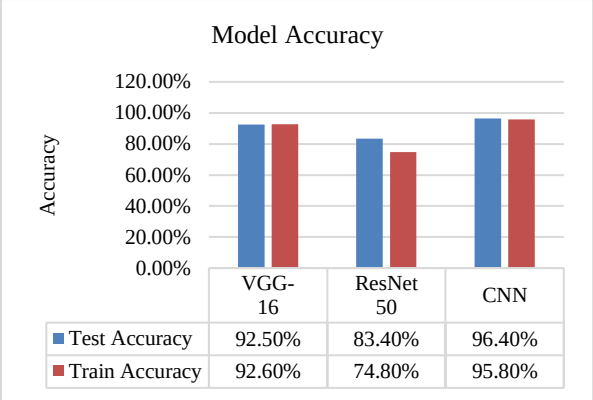

Fig. 6: Accuracy of the Deep Learning Model

The entire dataset is labeled and divided 75% for training, 20% for testing, and 5% for validation Using CNN as the trained model to train the proposed method's to training dataset with 30 epochs and a batch size of 64, this work achieved a Test Loss: 0.204 and Train Loss: 0.252 and Test Accuracy: 96.4 %, and Train Accuracy: 95.8 % in the fig 7.

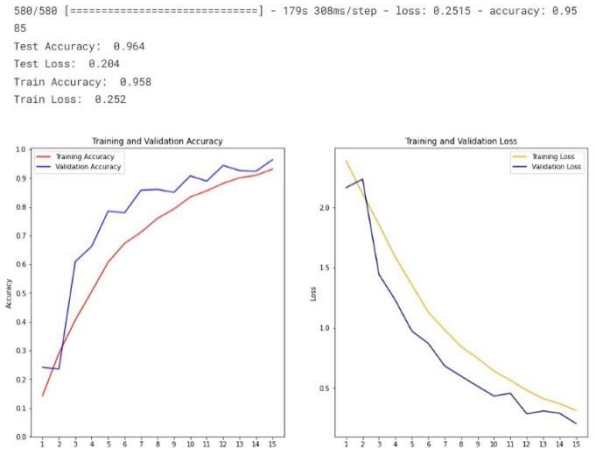

Fig. 7: Accuracy and Loss Evaluation for Training

4. CONCLUSION

Over the years, computer vision-based Human Action Recognition (HAR) systems have made significant strides, driven by the integration of advanced classification techniques like deep learning and machine learning. The success of HAR can be attributed to effective modeling, robust and discriminative representations, thorough analysis of dataset properties, and the deployment of fast and precise activity recognition systems.

This article introduces a deep learning-based method for human activity recognition, emphasizing the impact of the work custom dataset on activity classification and its potential for integration into automated computing systems. The proposed model demonstrates high accuracy, validated through extensive testing on diverse video samples. Human Activity Recognition (HAR) has garnered considerable attention and is applied across various domains, including video retrieval, monitoring, human-computer interaction (HCI), medical diagnostics, anomaly detection, surveillance, and suspicious behavior identification. Future research directions include expanding the range of activity classes in the dataset, exploring alternative perspectives, and designing recognition systems that are more efficient in terms of complexity and memory usage. Despite advancements, recognizing and interpreting human activities

remains challenging due to factors such as complex backgrounds, crowded scenes, and the need to derive activity patterns and constraints directly from raw video data.

5. REFERENCES

[1] Manaf, A. and Singh, S., 2021, May. Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications. In 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 110-114). IEEE.

[2] Vahora, S.A. and Chauhan, N.C., 2019. Deep neural network model for group activity recognition using contextual relationship. Engineering Science and Technology, an International Journal, 22(1), pp.47-54.

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

[4] Shikha, M., Kumar, R., Aggarwal, S. and Jain, S., 2020. Human activity recognition. International Journal of Innovative Technology and Exploring Engineering, 9(7), pp.903-905.

[5] Wei, L. and Shah, S.K., 2017, February. Human Activity Recognition using Deep Neural Network with Contextual Information. In VISIGRAPP (5: VISAPP) (pp. 34-43).

[6] Ravi, D., Wong, C., Lo, B. and Yang, G.Z., 2016, June. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. In 2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN) (pp. 71-76). IEEE.

[7] Hayat, A., Morgado-Dias, F., Bhuyan, B.P. and Tomar, R., 2022. Human Activity Recognition for Elderly People Using Machine and Deep Learning Approaches. Information, 13(6), p.275.

[8] Wu, D., Sharma, N. and Blumenstein, M., 2017, May. Recent advances in video-based human action recognition using deep learning: A review. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 2865-2872). IEEE.

[9] Khan, I.U., Afzal, S. and Lee, J.W., 2022. Human activity recognition via hybrid deep learning based model. Sensors, 22(1), p.323.

[10] Amrutha, C.V., Jyotsna, C. and Amudha, J., 2020, March. Deep learning approach for suspicious activity detection from surveillance video. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 335-339). IEEE.

[11] Jobanputra, C., Bavishi, J. and Doshi, N., 2019. Human activity recognition: A survey. Procedia Computer Science, 155, pp.698-703.

[12] Gu, F., Chung, M.H., Chignell, M., Valaee, S., Zhou, B. and Liu, X., 2021. A survey on deep learning for human activity recognition. ACM Computing Surveys (CSUR), 54(8), pp.1-34.

[13] Ann, O.C. and Theng, L.B., 2014, November. Human activity recognition: a review. In 2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014) (pp. 389-393). IEEE.

[14] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L. and Amirat, Y., 2015. Physical human activity recognition using wearable sensors. Sensors, 15(12), pp.31314-31338.

[15] Martínez-Mascorro, G.A., Abreu-Pederzini, J.R., Ortiz-Bayliss, J.C. and Terashima-Marín, H., 2020. Suspicious behavior detection on shoplifting cases for crime prevention by using 3D convolutional neural networks. arXiv preprint arXiv:2005.02142.

[16] Kaluza, B., 2013. Detection of Anomalous and Suspicious Behavior Patterns from Spatio-Temporal Agent Traces.

[17] Khare, S., Sarkar, S. and Totaro, M., 2020, June. Comparison of Sensor-Based Datasets for Human Activity Recognition in Wearable IoT. In 2020 IEEE 6th World Forum on Internet of Things (WF-IoT) (pp. 1-6). IEEE.

[18] Ahmad, Z., Illanko, K., Khan, N. and Androutsos, D., 2019, August. Human action recognition using convolutional neural network and depth sensor data. In Proceedings of the 2019 International Conference on Information Technology and Computer Communications (pp. 1-5).

[19] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L. and Amirat, Y., 2015. Physical human activity recognition using wearable sensors. Sensors, 15(12), pp.31314-31338.

[20] Subetha, T. and Chitrakala, S., 2016, February. A survey on human activity recognition from videos. In 2016 international conference on information communication and embedded systems (ICICES) (pp. 1-7). IEEE.

[21] Jegham, I., Khalifa, A.B., Alouani, I. and Mahjoub, M.A., 2020. Vision-based human action recognition: An overview and real world challenges. Forensic Science International: Digital Investigation, 32, p.200901.

[22] Sarkar, P. K., & Abdullah, A. B. M. (2022). Diagnosing Suspects by Analyzing Human Behavior to Prevent Crime by Using Deep and Machine Learning.

[23] Arunnehru, J., Chamundeeswari, G. and Bharathi, S.P., 2018. Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos. Procedia computer science, 133, pp.471-477.

[24] Varshney, P., Harsh Tyagi, N.K., Lohia, A.K. and Girdhar, P., 2021. A Deep Learning Based Approach to Detect Suspicious Weapons. Proceedings http://ceur-ws. org ISSN, 1613, p.0073.

[25] Khattar, L., Kapoor, C. and Aggarwal, G., 2021, January. Analysis of Human Activity Recognition using Deep Learning. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 100-104). IEEE