VISION-BASED HUMAN ACTIVITY RECOGNITION USING DEEP LEARNING TECHNIQUES

Divye Gupta[a*], Aarnav JP[b*], Mayank Hora[c], Muskaan Bhardwaj[d], Rachna Narula[e*]

Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, 110063, New Delhi

a* divyegupta00@gmail.com , b aarnavjp@gmail.com , c mayankhora8@gmail.com , d bhardwajmuskaan19@gmail.com, e* rachna.narula@bharatividyapeeth.edu

ABSTRACT

Technology's growing use has facilitated the quality of living. It can provide Improved Security and Monitoring, Enhanced Assistance for Daily Living, Advancements in Healthcare, and Smarter Homes and Environments. Artificial Intelligence (AI) is the field that aims to define how human intelligence is mimicked by machines which are programmed to think or behave like humans. Artificial intelligence (AI) is the intelligence that humans have implanted in machines, as opposed to natural intelligence. Artificial intelligence research aims to simulate human intellect in robots. This can be accomplished in a number of ways, such as deep learning—a kind of machine learning that draws inspiration from the human brain—and machine learning, which teaches computers from data without the need for explicit programming. Deep learning is that domain of artificial intelligence that allows the evolution of intelligent software and machines ideally suited to assist people with all of their activities of daily life. Advances in natural language processing through deep learning allow software to comprehend and react to human language in ways never possible before. This encourages the creation of virtual assistants that comprehend our requests, chatbots that can carry on conversations, and machine translation programmes that can translate between different languages. Recent developments in the domains of artificial intelligence (AI) and machine learning (ML) made it feasible to develop new methods and instruments for assessing human behaviour. Due to its applicability in a number of industries, comprising entertainment, security and surveillance, health, and intelligent environments, human activity recognition has gained prominence significantly. Human activity recognition (HAR) using video sensors typically involves analyzing the visual data captured by cameras to classify and identify the actions of individuals. In the following paper, we propose ConvLSTM and LRCN-based HAR. A huge variety of films from the publicly accessible data set, UCF50 comprising a wide range of activity classes that are used to build a statistical model. For the model proposed in this paper, the accuracy has turned out to be 94%, the average f1-score is 0.93 and the average recall is calculated to be 0.925. The Loss curve has also been plotted along with the accuracy curve for the proposed model for recognizing human activities.

Keywords: Human Activity Recognition, Deep Learning, Neural Networks, LSTM, ConvLSTM, LRCN.

1. INTRODUCTION

Human Activity Recognition (HAR) deals with automated human physical activity identification [1]. The primary objective of HAR is to recognize any unusual activity or incident and analyze human activities [14]. Unlike most species, humans are distinguished by their ability to perform different activities simultaneously and their ability to learn new tasks and teach their offspring how to perform them. These activities can range from as simple as walking to as complex as a pommel horse [16].

HAR is used in various industries, including healthcare, entertainment, education, and security. It has been gaining a lot of focus among researchers because of its multitude of applications like video analytics, border infiltration detection, computer vision, biometrics, digital libraries, video surveillance, and many more [17, 14, 15].

HAR systems are built and categorized into mainly three types of sensors - 1) vision sensors, consisting of 3d cameras, depth cameras, skeleton analysis, etc., 2) environmental sensors, including device (smartphones) and wearable sensors, and 3) radar sensors, which are the device-free wireless sensors. Deep learning (DL) has proven to be a good approach for understanding HAR systems, but choosing the appropriate DL method can prove to be a bit challenging task [1, 22].

Video-based HAR is useful in structuring the video monitoring system that helps to identify unusual events. The system is based upon detecting the accurate activity with the help of behavior shown in the frame sequence of a particular video, without manual operations [15]. Currently, there is a lot of content available in the form of video footage on the Internet, and it is impossible to cater to the demands of accurately identifying and annotating all the videos and extracting features based on human labor. Thus, there is a need for an automated system. The process can be described in 3 parts: 1. Framing - monitoring the behavior of an individual by extracting frames from the video of a given dataset, 2. Feature processing - processing of data with the purpose of fetching particular characteristics, 3. Classification - detecting human activity through the model designed to obtain the results from previously obtained features [14].

Despite extensive research that has been done in the field of HAR , it is still taxing because of the limitations like disturbance from the background, frequent change in brightness, relative motion of the object, object occlusion, etc. [17]. Due to the intricacy of the tasks, the caliber of the data being processed, high dimensionality, intraclass variability, and interclass similarity, it may prove to be a difficult task [16].

Deep learning offers a wide area for working on HAR. Applying deep learning to HAR has significantly improved the accuracy of action detection by overcoming several challenges that classical machine learning techniques encountered. In order to learn effective classifiers from unstructured data, it offers a data-driven methodology that establishes a hierarchy between low-level qualities and high-level abstractions. Deep learning excels in automatically extracting characteristics that are appropriate for the job at hand. Avoid relying on heuristic hand-crafted functions and scale to more complex behaviour detection tasks [18].

Further, deep learning has an important value in implementing self-learning as well as transfer learning [7]. Convolutional neural networks (CNN), which are useful for extracting spatial information, are used in certain deep learning systems, while others advice employing long short-term memories (LSTM), which are rich in temporal information. Deep feed-forward neural networks and their variations are used by some approaches in this domain. Each of these approaches has its own pros and cons [9]. However, we found that the strength of one technique can be exploited to improve the robustness of the other.

Most of the earlier work considered HAR as an individual classification method. A HAR model will typically extract pre-built features from a sequence of activity data and link the features to activity labels. The most prevalent statistical and structural features are those that are manually constructed. These properties are utilized often in many time-series issues, although they are heuristic and not task-dependent. Deep learning methods are currently being used by some researchers to look for deep HAR features. The process of acquiring data representations is the foundation of the "deep learning" family of machine learning algorithms. a popular deep learning technique.

One advantage of Long Short-Term Memory (LSTM) networks, Recurrent neural networks (RNNs), and Convolutional neural networks (CNNs) is their ability to automatically identify deep characteristics. Unlike CNN, which is a feedforward neural network, RNN displays dynamic temporal behaviour by adding directed cycles. Stated differently, RNNs can learn temporal correlations by utilising "memory" and time series data. LSTM networks can effectively address the problem of long-term dependency on RNNs by adding more complex memory cells to RNNs.

In the subsequent study, we put forth a model constructed using an LSTM and CNN combination.
By combining the strengths of both Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), we can extract temporal information across activities and utilise CNN's strength in feature extraction [1]. This work aims to develop a more sophisticated

artificial intelligence that can recognise everyday activities (ADLs) including walking, jogging, and running in real time.[20]

We have conducted experiments using the UCF-50 dataset, which includes 50 daily exercise courses including basketball, diving, horse racing, kayaking, push-ups, swings, walking with dogs, etc., to check the effectiveness and resilience of our suggested methodology. Compared to other vision-based methods for recognising human activities, accuracy has been determined to be the most important evaluation metric.

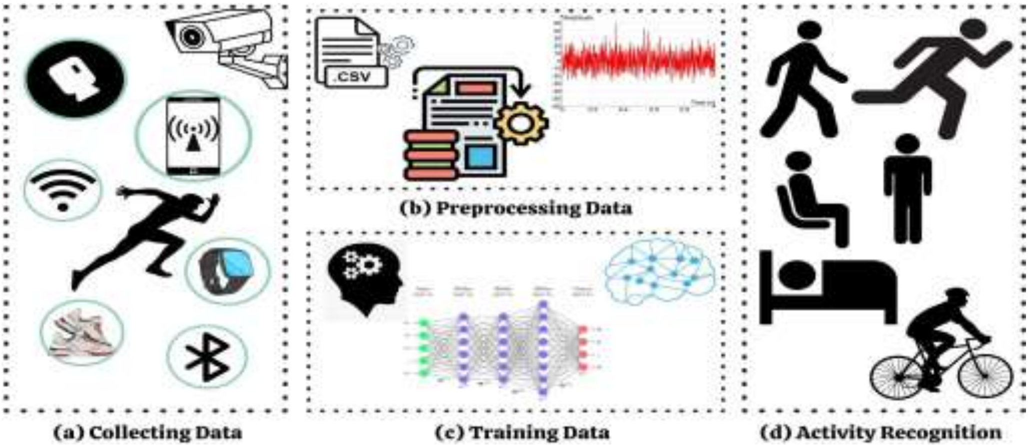


Fig 1.1: Basic Structure of HAR

A typical mechanized system of HAR consists of data acquisition, activity detection, modelling, and finally, classification. The Human Activity Recognition framework comprises four major sections,

a) data collection for a dataset based on vision;

b) data pre-processing, which carries out crucial preprocessing operations on the gathered data;

c) training of the data, that uses of machine learning (ML) or deep learning (DL) methodology to learn patterns from the gathered data; and

d) recognition of the performed actions.

2. LITERATURE REVIEW

Human Activity Recognition has become a widely researched topic within the field of computer vision. It is also a very challenging problem of time series categorization that involves the prediction of an individual's movement and actions utilising a series of data collected by sensors. HAR is traditionally based on deep learning techniques and along with that it requires signal processing as well as processes to carefully construct features out of the raw data collected to suit the best machine learning model.

Over the past years, there has been a rapid advancement in technology and various Human activity recognition models have been created for aid in the automation of visual monitoring systems. Numerous surveys have been done which summarise various research work done on several different methodologies in the field of action recognition. A paper by

Zawar Hussain [4] very elaborately discusses various categorizations of techniques used in the HAR models. The data collected by sensors for action recognition can generally be classified into three main sensor approaches namely Vision based sensors, Radio-Frequency sensors, and Environment sensors. Zawar Hussain [4] thoroughly describes various sensor-based approaches in the field of action detection and also focuses on device-free human activity recognition categories along with the various comparison metrics to evaluate the created model on the accuracy, technology, loss, approach, cost, latency, and much more.

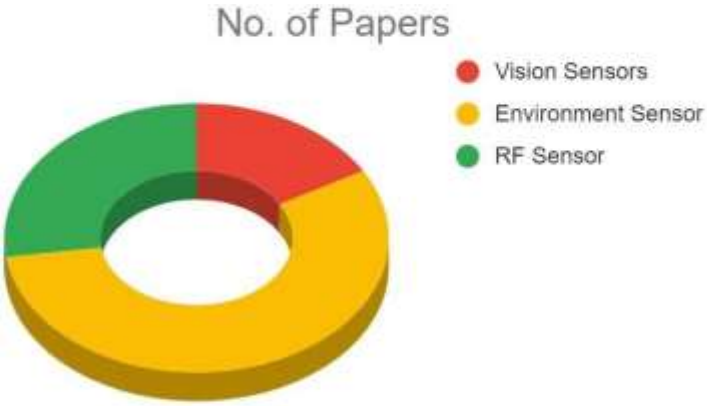


Fig 2.1: Classification of HAR

Vision-based sensors rely heavily on the ability to see their surroundings. Device sensors and wearable sensors are part of environmental sensors.
The first approach involves attaching sensors to commonly used objects to recognize human actions based on the user's interaction with both the environment and these objects. In contrast, the second approach directly interfaces with the human body and utilizes a variety of sensors, such as an accelerometer, gyroscope, and magnetometer. Basically, device-free sensors and radio-frequency sensors pick up signals that are influenced by human activity in the channel.

According to studies on various Human Activity Recognition models **[Fig 2.1]**, 16.7% of the models employed vision sensors, 27.3% utilized radio frequency sensors, and the remaining 56.1%—a substantial portion of Human Activity Recognition models—made use of environment sensors.

In this paper, we use the UCF-50 data set, a collection of 50 action categories that altogether comprises 6676 realistic videos. A citation published by Shian-Ru Ke [2] is an excellent resource that undertakes a thorough survey of the vision-based models for Human Activity Recognition. Shian-Ru Ke [2] discusses the numerous domains of applications of HAR models along with the three crucial phases of processing which include activity detection and classification techniques, human object segmentation, feature extraction, and representation. Three aspects of human action recognition are discussed: foundational technology, human action detection frameworks, and representation of applications from lower-levels to higher-levels.

Preksha Prateek [11] provides an overview of the methods currently in use for HAR on trimmed videos and discusses an action recognition task's comprehensive framework which entails feature extraction, feature encoding, dimensionality reduction, action classification, and various action classification methods, as well as the beneficial and adverse characteristics of the mentioned tasks, along with readily accessible data sets.

The computational cost of the processing of any video is quite high. An appropriate frame selection and extraction are essential parts of video processing for activity recognition. Shreyank N Gowda [5] showed that proper frame selection enhances the model’s functionality and accuracy even in the domain of trimmed video and a method was proposed for better frame selection in which the best frames are very efficiently spread across the video. In this method, a large number of frames from a video are considered jointly rather than selecting one frame at a time. A two-fold framework by Soumya Ranjan Mishra [6] helps to reduce the computational load of regular frame extraction by extracting the frames only when the action is committed.

Juhi Singh [3] asserts that when compared to conventional CCTV motion detection systems, vision-based HAR techniques can provide an end-to-end automated home monitoring system with an accuracy of more than 93%. We can increase the system's accuracy and get notably better results by employing the LRCN (Long Term Recurrent Convolutional Network) method. According to Zeqi Yu and Wei Qi Yan [7], background noises, changes in perspective, and other complexities, the prediction of the model is affected. To deal with these issues, three algorithms were designed, namely, two-stream CNN, CNN+LSTM, and 3D CNN.

Ronald Mutegeki [9] states that CNN and LSTM both have been the subject of in-depth inquiry in the past in isolation, therefore presenting a CNN-LSTM classifier that It improves the accuracy of identifying the action performed while reducing the model's complexity. Combining CNN with LSTM has the potential to both simplify the model and increase the forecast accuracy. The aim of Analysis of Human Activity Recognition (HAR) is identification of the activity a device's user is carrying out by employing sensors like the accelerometer, gyroscope, magnetometer, and others that are incorporated into IMU devices and smartphones. Kai Hu [12] stated that due to the ability to extract time information, Long Short-Term Memory. The LSTM network serves as a conventional method for activity detection. It's recommended to improve the input differential feature module and extend the network to incorporate a spatial memory state differential module to create an enhanced Spatio-Temporal Differential Long Short-Term Memory (ST-D LSTM) network. Furthermore, a transmission mode for ST-D LSTM is proposed, enabling horizontal transfer of the spatial memory state via ST-D LSTM units. To verify the effectiveness of the new network, these enhancements are applied to traditional Long-term Recurrent Convolutional Networks (LRCN).

A hybrid architecture highlighting the features of both Convolutional Neural Networks (CNN) as well as Long-Short Term Memory (LSTM) was also presented by Samundra Deep and Xi Zheng [8] which was tested on the UCI-HAR dataset comprising sensor data collected through an accelerometer and gyroscope in a smartphone. Samundra Deep [8] performed the same experiment on the LSTM model and the accuracy of the CNN-LSTM model was better than the other one. The combined average results from several classifiers tuned their performances and a generic activity recognition framework was provided in a model proposed by Vijay Bhaskar Semwal [10] as an ensemble learning implementation. In our current research, we propose a model for action recognition employing the UCF-50 dataset and two algorithms: the Long-term Recurrent Convolutional Network (LRCN) along with a Convolutional Neural Network (CNN) combined with Long-Short Term Memory (LSTM).

3. DATASET

The training of a Human Action Recognition model demands a substantial collection of videos of people doing various actions. In this paper, we have used the UCF50 action recognition data set which is a collection of 50 activity classes that offers a realistic set of data for recognizing activities obtained from YouTube videos. Dataset can be found at https://www.crcv.ucf.edu/data/UCF50.php It also offers a diverse collection of human activities, given the variety of camera motion, poses, object appearances, viewpoints, cluttered backgrounds, and various illumination in the surroundings. The videos in the same activity class may share some features like the same person performing different activities. Every single video out of the 6618 available video data set belongs to one of the fifty activity classes. The entire set of data is segregated into train and test data having 75% and 25% of the total videos respectively but prior to that, the data set is shuffled to decrease bias and generate splits that precisely depict the distribution of the data overall.

4. PROPOSED METHODOLOGY

Data is first collected using strategically placed video sensors. These sensors could be either stationary cameras monitoring a certain region or mobile cameras that people wear on their person. After it is obtained, the video data is pre-processed to prepare it for analysis. This could mean making adjustments for variations in lighting, maintaining consistent video frame sizes, or setting the focal point out from the background. Feature extraction is HAR's main purpose. Here, essential elements that faithfully capture human behaviours are extracted from the pre-processed video frames. Two popular techniques are pose estimation and optical flow. Pose estimation tracks and locates critical joints in each frame, like the elbows, knees and ankles. This enables the system to gradually learn the user's posture and movement habits. On the other hand, optical flow examines the pixel motion patterns across successive frames to capture the dynamics of an activity, including movement direction and speed.

After feature extraction, the main focus shifts to a classification model. This model was previously trained on an extensive dataset of tagged video examples exhibiting a variety of human behaviours. It is usually powered by machine learning technologies such as deep learning. Every potential action category is given a likelihood score by the classification model, which is based on the attributes that were collected from the newly released video clip. Popular choices for classification models include Convolutional Neural Networks (CNNs) and

Recurrent Neural Networks (RNNs), which can manage the sequential nature of video data and identify patterns in visual data, respectively.
The last step is to analyse the results that the categorization model has produced. Every action category in the dataset has a probability score assigned by the model. The system identifies the particular action being carried out in the video footage based on the category with the highest score.

4.1 Human Activity Recognition with Convolutional Neural Networks

CNNs are widely used models in deep learning computer vision. The visual cortex of the human brain and CNN have comparable architectures. Filters can be applied to input images to detect objects and extract spatial or temporal features. Certain completely linked layers are utilized for classification, and convolution layers are made up of filters. CNNs can learn features and handle large datasets with ease because to pooling layers. The primary goal of pooling is, in fact, to minimize the dimensionality of incoming data and extract the most common characteristics.[1]
An additional inquiry by [22] CNN is a network that has multiple layers to its construction. The design consists of two basic components:
1) a network with connections, and
2) several layers for convolution and sampling.
The latter is used in feature extraction, whereas the former is required for studying classification weights.

A normal CNN consists of three layers:
1. Convolutional layer;
2. Pooling layer; and
3. Fully-connected layer.
The convolutional layer's filters, sometimes referred to as the feature map, help extract local features. One filter is used to create a single feature map. To create a feature map, you have to slide filters over the input data to create the dot product. We call this procedure a convolution operation. The receptive field, a small portion of the input data the same size as the filter, is connected to every single neuron in the feature map. The neuronal weightings of a feature map are made publically available, significantly reducing the number of parameters needed to provide meaningful results. Furthermore, it becomes possible to identify and recognize specific patterns regardless of where they appear in the input. The stride number and filter size are used to calculate the magnitude of the feature map.

4.2 Human Activity Recognition with Conv. Long Short-Term Memory

It is believed that LSTM, a time-based network of recurrent neurons, is suitable for time series data [21]. Horizontal lines traverse the top of the graph as the LSTM assesses the condition of the cells. The cell state is easily contrasted and compared to a conveyor belt, and it may be applied across the entire chain with just a few straightforward linear exchanges.[21] The LSTM cells continue to carry this information. It also provides the ability to add or remove information from the cell's state using a sophisticated architecture called a "gate." The gate is made up of the sigmoid layer of a neural network and a pointwise multiplication operation, just as a door provides a means of passing with knowledge.The distinguishing characteristic of LSTM is its ability to modify the self-loop weight without influencing gradient expansion or vanishing gradients, even with fixed model parameters. This is achieved by increasing the input, forgetting, and output thresholds. This makes it possible to dynamically alter the integral scale at various times. LSTM is a technology that has multiple applications. They are widely used in many different industries, including handwriting recognition, robotics, translation, and image analysis.

4.3 Human Activity Recognition with Long-term Recurrent Convolutional Network (LRCN)

The LRCN approach combines LSTM and convolution layers into a single framework. Another sensible alternative is to combine a CNN model with an independently trained LSTM model. Using CNN and a trained model, it is possible to extract spatial information from video frames and optimize this model for the intended use. As a result, the LSTM prototype may be able to forecast actions in the video with the aid of data obtained by the CNN model. In contrast, the Long-term Recurrent Convolutional Network (LRCN) integrates the CNN and LSTM layers into a single model. At each time step, the gathered spatial data is fed into an LSTM layer for temporal sequence modeling with convolutional layers. By using this method, the network learns spatiotemporal features through an end-to-end training session, resulting in the production of a stable model. A Long-term Recurrent Convolutional Network (LRCN) model is useful for applications requiring both language and visual inputs and outputs. This approach combines a deep hierarchical visual feature extractor with CNN-like recognition and synthesis of temporal dynamics.[13]

5. PROPOSED MODEL

The model presented in this paper uses two methodologies namely, ConvLSTM and LRCN. Initially, the UCF-50 data set, comprising 50 activity classes, is preprocessed for data training. The frames of the videos are first resized, normalized, and then extracted. In addition to that the feature extraction is done on the extracted frames. The dataset is then shuffled and split into train i.e., 75%, and test i.e., 25% of the total videos. LSTM layers are employed for temporal detection, whereas CNN aids in spatial extraction of the frames. ConvLSTM cells are LSTM network variants that include convolutional processes. It is an LSTM with built-in convolution, which enables it to distinguish between spatial input components and take into consideration the temporal relationship. As a result, the convolutional LSTM is capable of taking in 3d input whereas, LSTM solely could take in only 1d input. The number of epochs taken into account is 50, with the initial number of filters and batch size equal to 4 each. With every fusion, the number of filters keeps on increasing while with every pooling, the size of filters keeps on decreasing. Time distributed layer is used at a 20% dropout rate. A softmax classifier is used for classification. For the purpose of optimization, Adam optimizer is used. The ConvLSTM2D layer's output is flattened before being given to the tanH activation function and dense layers. The MaxPooling3D and Dropout layers are employed to minimise frame dimensions and avoid overfitting of the model.

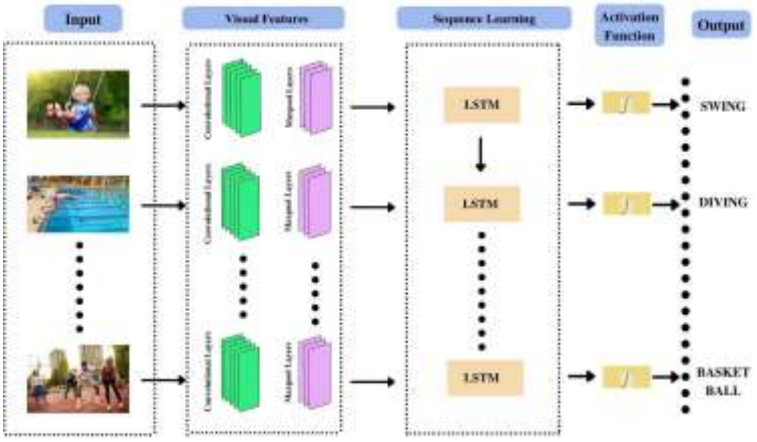
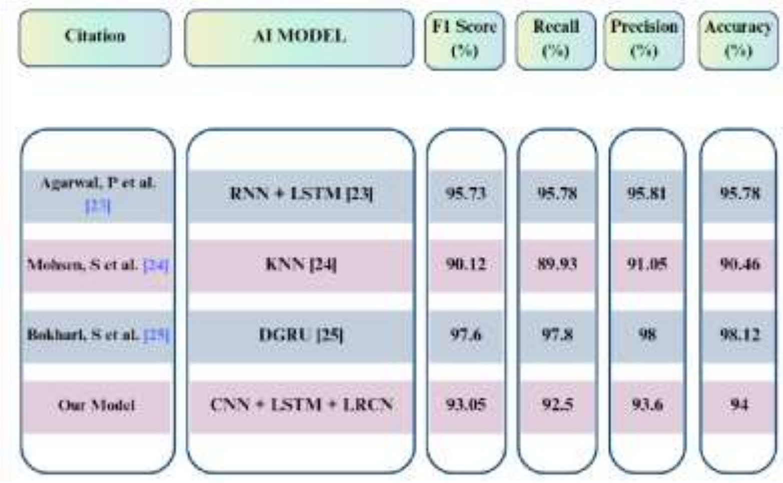


Fig 5.1: Structure of proposed model (convLSTM)

6. DISCUSSIONS

The following figure **[Fig 6.1]** presents a comparative study of six different types of AI models, each with their respective F1 scores, Recall, Precision, and Accuracy metrics. The authors of these models are also listed in the table. Our model's performance is also included:

F1 score of 93.05%, Recall of 92.5%, Precision of 93.6%, and Accuracy of 94%. The results of the comparison highlight the effectiveness of the models in comparison to each other, with our model outperforming the KNN model in several key metrics. The table provides valuable insight into the state of AI models and their performance, allowing researchers to make informed decisions when choosing the most appropriate model for their specific needs.



Citation	AI MODEL	F1 Score (%)	Recall (%)	Precision (%)	Accuracy (%)
Agarwal, P et al. [23]	RNN + LSTM [23]	95.73	95.78	95.81	95.78
Mohsen, S et al. [24]	KNN [24]	90.12	89.93	91.05	90.46
Bokhari, S et al. [25]	DGRU [25]	97.6	97.8	98	98.12
Our Model	CNN + LSTM + LRCN	93.05	92.5	93.6	94

Fig 6.1: Comparative study of various models

7. RESULTS

Our proposed model is studied using various evaluation metrics. For some of the activities from the specified 50 classes of the UCF-50 dataset, recall, f1 scores, and precision have been calculated in order to identify the positive predictive value and the sensitivity.

Activity No. (Name)	Precision	Recall	F1-score
0 (Swing)	0.5	1	0.67
2 (Horse Race)	1	1	1
3 (Diving)	1	1	1
4 (Basketball)	1	0.50	1

Table 7.1: Performance key metrics of some activities

Visual results showing the values of activities are given in **[Table 7.1]**. For attaining more clarity on the values, we have considered only 4 activities, present on the leftmost column of the table, namely Swing (denoted by 0), Horse Race (denoted by 2), Diving (denoted by 3), and lastly, Basketball (denoted by 4).

As shown in the table, the precision of all the activities on average is around 87.5%. The f1-score for the activities - Swing, Horse race, Diving, and Basketball is 0.67, 1.00, 1.00, and 1.00 respectively, while the recall is 1.00, 1.00, 1.00, and 0.50 respectively. The magnitude of these values proves the effectiveness of our proposed technique. The class-wise performance for each activity present in the dataset can be evaluated in a similar manner. The accuracy of the model is 94%. It is safe to say that the convLSTM approach works well for vision-based human activity recognition.

Key Metrics	Performance
F1 Score	93.05%
Recall	92.5%
Precision	93.6%
Accuracy	94%

Table 7.2: Overall performance key metrics of the model

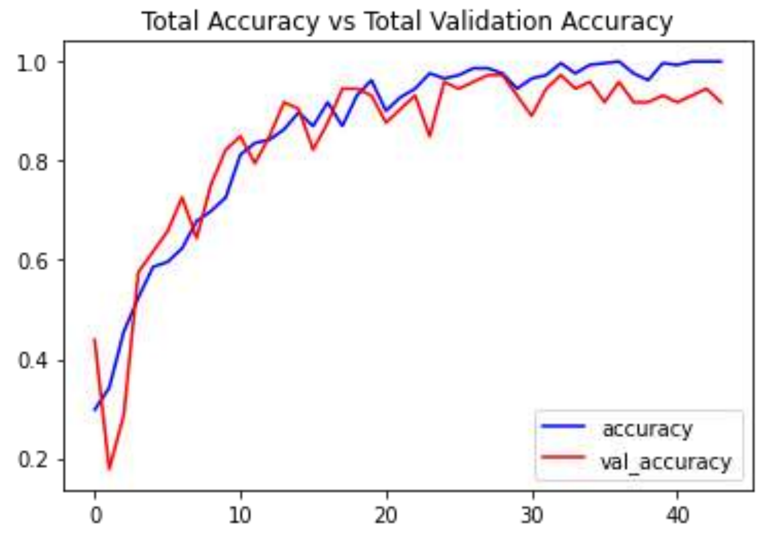


Fig 7.1: Total Accuracy vs Total Validation Accuracy curve

The above graph **[Fig 7.1]** shows the comparison between the testing accuracy and the validation accuracy. From the graph, we can conclude that after the complete training, the mode accuracy does not decrease by a lot for validation or new data that is fed into the model. This shows us that the model is quite robust and is not overfitting in any manner.

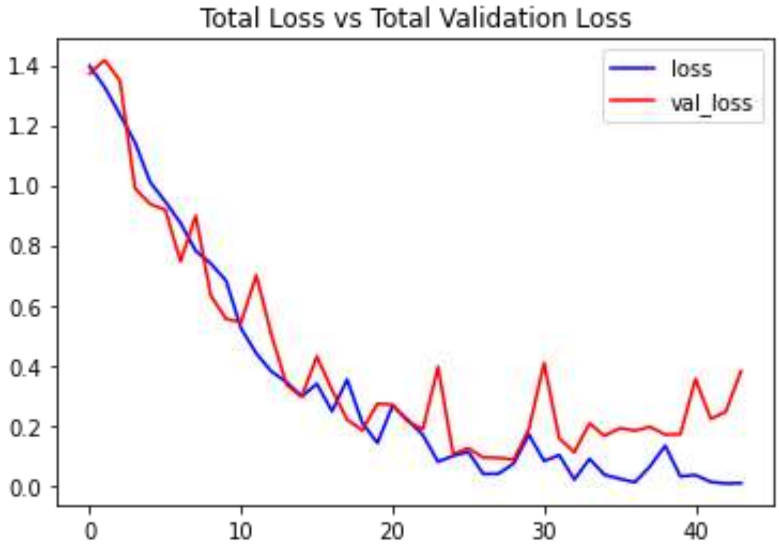


Fig 7.2: Total Loss vs Total Validation Loss curve

From the above graph **[Fig 7.2]**, we are able to conclude that the false positives of our model are quite low as the losses decrease gradually. And the losses for the validation data are also very comparable to the total losses for the testing data.

8. CONCLUSION

We have presented a Convolutional LSTM network approach to cater to the ever-growing domain of Activity Prediction in Humans. Identification of the spatial region in the individual frames of the videos of different activity classes and the temporal effect across those frames was carried out effectively. The activities are classified with the help of these convLSTM cells and the name of the activity is measured as an output for checking the accuracy. This approach is capable of giving more efficient results by consuming less time and expense. Loss and accuracy curves were also plotted to obtain a better perception of the efficiency of the model. The performance measures, recall, f1-score, and precision also helped to demonstrate the validity of the technique. A further benefit is an idea of employing a Long-term Recurrent Convolutional Network (LRCN) technique. The spatial information from the frames is extracted using CNN layers, and the LSTM layers are then fed to the individual frames one at a time to simulate the temporal sequence. This spatio-temporal feature technique results in a robust model.

9. REFERENCES

1. Abbaspour, S.; Fotouhi, F.; Sedaghatbaf, A.; Fotouhi, H.; Vahabi, M.; Linden, M. A Comparative Analysis of Hybrid Deep Learning Models for Human Activity Recognition. *Sensors* 2020, *20*, 5707.
2. Ke, S. R., Thuc, H. L. U., Lee, Y. J., Hwang, J. N., Yoo, J. H., & Choi, K. H. (2013). A review on video-based human activity recognition. *Computers*, *2*(2), 88-131.
3. Singh, J., & Sinha, S. Video Based Human Activity Recognition Surveillance System.
4. Hussain, Z., Sheng, Q. Z., & Zhang, W. E. (2020). A review and categorization of techniques on device-free human activity recognition. *Journal of Network and Computer Applications*, *167*, 102738.
5. Gowda, S. N., Rohrbach, M., & Sevilla-Lara, L. (2021, May). Smart frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 2, pp. 1451-1459).
6. Mishra, S. R., Mishra, T. K., Sanyal, G., Sarkar, A., & Satapathy, S. C. (2020). Real time human action recognition using triggered frame extraction and a typical CNN heuristic. *Pattern Recognition Letters*, *135*, 329-336.
7. Yu, Z., & Yan, W. Q. (2020, November). Human action recognition using deep learning methods. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.
8. Deep, S., & Zheng, X. (2019, December). Hybrid model featuring CNN and LSTM architecture for human activity recognition on smartphone sensor data. In *2019 20th international conference on parallel and distributed computing, applications and technologies (PDCAT)* (pp. 259-264). IEEE.
9. Mutegeki, R., & Han, D. S. (2020, February). A CNN-LSTM approach to human activity recognition. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 362-366). IEEE.

10. Semwal, V. B., Gupta, A., & Lalwani, P. (2021). An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition. *The Journal of Supercomputing*, *77*(11), 12256-12279.
11. Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259-2322.
12. Hu, K., Zheng, F., Weng, L., Ding, Y., & Jin, J. (2021). Action Recognition
13. Algorithm of Spatio–Temporal Differential LSTM Based on Feature
    a. Enhancement. *Applied Sciences*, *11*(17), 7876.
14. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
15. Sansano, E., Montoliu, R., & Belmonte Fernandez, O. (2020). A study of deep neural networks for human activity recognition. *Computational Intelligence*, *36*(3), 1113-1139.
16. Tasnim, N., Islam, M. K., & Baek, J. H. (2021). Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. *Applied Sciences*, *11*(6), 2675.
17. Hernández, F., Suárez, L. F., Villamizar, J., & Altuve, M. (2019, April). Human activity recognition on smartphones using a bidirectional LSTM network. In *2019 XXII symposium on image, signal processing and artificial vision (STSIVA)* (pp. 1-5). IEEE.
18. Kushwaha, A., Khare, A., & Srivastava, P. (2021). On integration of multiple features for human activity recognition in video sequences. *Multimedia Tools and Applications*, *80*(21), 32511-32538.
19. Murad, A., & Pyun, J. Y. (2017). Deep recurrent neural networks for human activity recognition. *Sensors*, *17*(11), 2556.
20. Peng, L., Chen, L., Ye, Z., & Zhang, Y. (2018). Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1-16.
21. Almaslukh, B., Artoli, A. M., & Al-Muhtadi, J. (2018). A robust deep learning approach for position-independent smartphone-based human activity recognition. *Sensors*, *18*(11), 3726.
22. Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, *25*(2), 743-755.
23. Mekruksavanich, S., & Jitpattanakul, A. (2021). Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. *Electronics*, *10*(3), 308.
24. Agarwal, P., & Alam, M. (2020). A lightweight deep learning model for human activity recognition on edge devices. *Procedia Computer Science*, *167*, 2364-2373.
25. Mohsen, S., Elkaseer, A., & Scholz, S. G. (2022). Human activity recognition using K-nearest neighbor machine learning algorithm. In *Sustainable Design and Manufacturing: Proceedings of the 8th International Conference on Sustainable Design and Manufacturing (KES-SDM 2021)* (pp. 304-313). Springer Singapore.
26. Bokhari, S. M., Sohaib, S., Khan, A. R., & Shafi, M. (2021). DGRU based human activity recognition using channel state information. *Measurement*, *167*, 108245.
27. Mekruksavanich, S., & Jitpattanakul, A. (2020, October). Smartwatch-based human activity recognition using hybrid lstm network. In *2020 IEEE SENSORS* (pp. 1-4). IEEE.
28. Malki, Z., Atlam, E., Dagnew, G., Alzighaibi, A. R., Ghada, E., & Gad, I. (2020). Bidirectional residual LSTM-based human activity recognition. *Computer and Information Science*, *13*(3), 40.