# *Human Activity Recognition Using Conv-LSTM with Infrared Camera Surveillance*

Bindu Madhavi Tummala
*Department of Computer Science and Engineering*
*Velagapudi Ramakrishna Siddhartha Engineering College*
Vijayawada, Andhra Pradesh, India
bindumadhavi@vrsiddhartha.ac.in

Chitturi Purna Chandra Sekhar
*Department of Computer Science and Engineering*
*Velagapudi Ramakrishna Siddhartha Engineering College*
Vijayawada, Andhra Pradesh, India
chpchsekhar365@gmail.com

Andra Karthik
*Department of Computer Science and Engineering*
*Velagapudi Ramakrishna Siddhartha Engineering College*
Vijayawada, Andhra Pradesh, India
karthik.andra7@gmail.com

Anupam Gupta
*Department of Computer Science and Engineering*
*Velagapudi Ramakrishna Siddhartha Engineering College*
Vijayawada, Andhra Pradesh, India
anupamlast@gmail.com

***Abstract*—Human Activity Recognition (HAR) the usage of infrared (IR) video statistics leverages superior computer vision and deep gaining knowledge of techniques to detect and classify human activities in numerous environments, even beneath low or no-light situations. This is important for packages including safety, surveillance, and healthcare diagnostics. Deep mastering models, which includes convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, play a substantial position in this technique. CNNs extract spatial features from person frames, while LSTMs model time-structured video sequences, permitting the accurate popularity of complex activities. In this have a look at, the Conv-LSTM version demonstrated superior performance, accomplishing an accuracy of ninety seven.71% on an infrared video dataset. This method outperformed other trendy models, consisting of Enhanced YOLOv5, Attention-primarily based LSTM, Deep Conv-LSTM, and movement recognition models based totally on thermal motion pictures. These outcomes spotlight the effectiveness of the Conv-LSTM framework in infrared-primarily based pastime recognition, showcasing its capacity for real-time analytics in hard environments.**

***Index Terms*—Human Activity Recognition (HAR), Infrared Video, Deep Learning, Convolutional Neural Networks (CNNs), Long-Short-Term Memory (LSTM), and Surveillance.**

## I. Introduction

In today's safety-conscious environment, monitoring and analyzing human behavior has become critical for maintaining public safety and preventing security breaches. Traditional surveillance techniques often depend on manual monitoring by security personnel, which may be inefficient, time-consuming, and prone to human errors. As a result, critical incidents or potential threats may go unnoticed, leading to delayed responses and possibly dangerous situations.

Human Activity Recognition (HAR) technology offers a more reliable and efficient solution by automatically detecting, analyzing, and interpreting human movements in real-time. By leveraging advanced machine learning and deep learning models, HAR systems can recognize specific behaviors, such as abnormal or [1] suspicious actions, that may indicate potential security risks. These systems have the potential to distinguish between normal and unusual activities, enabling security personnel to focus on incidents that truly require attention.

Integrating HAR[4] systems into surveillance infrastructures, such as CCTV networks for infrared conditions, can significantly enhance situational awareness. In environments like airports, schools, public spaces, and other high-security areas, HAR technology can help prevent incidents by providing an early warning of potentially suspicious behaviors or emergencies. For example, it can detect behaviors such as loitering, unauthorized access[6] , or even physical confrontations, prompting a swift response from security teams.

Moreover, real-time activity recognition helps optimize resource use by alerting security personnel to critical situations immediately, rather than relying on hours of manual video analysis. This integration improves the efficiency of surveillance and ensures quicker responses to incidents in [9] infrared environmental conditions.,

## II. Literature Survey

Ujwalla Gawande et al. [1] proposed a framework aimed at improving human recognition and identifying suspicious activities in surveillance systems, especially in educational settings. They enhanced the YOLOv5 model with optical flow-based motion feature extraction, achieving a detection accuracy of 91.12%. The model's complexity, however, could limit its scalability in real-time applications.

Sunil Malviya et al. [2] introduced a framework for real-time detection of suspicious activities in crowded environments using motion impact maps. This approach achieved an accuracy of 90.59% and a recall of 90.42%.

Jae-hook Jeong et al. [3] developed an anomaly detection system that combined deep learning models, including a 3D Convolutional AutoEncoder and SlowFast neural network.

979-8-3315-4389-1/25/$31.00 ©2025 IEEE

This system achieved an accuracy of 85% but faced challenges due to its reliance on virtual datasets and computational complexity.

Manoj Kumar et al. [4] developed a deep learning algorithm using CNNs, Bi-LSTM, and attention mechanisms for abnormal human behavior detection. The system achieved an accuracy of 88.9% but struggled with complex datasets and high computational demands.

Abid Mehmood et al. [5] proposed a two-stream CNN algorithm for real-time crowd anomaly detection, achieving detection accuracy of 91.12%. Despite its reduced computational cost, it faced challenges in adapting to different real-world situations.

Jaouedi et al. [6] introduced a hybrid deep learning algorithm combining Gated Recurrent Neural Networks (GRNN) with Gaussian Mixture Models (GMM) and Kalman filters. The system achieved an average accuracy of 87.3% on the KTH dataset but faced challenges in scalability due to its reliance on high-quality input data.

Wang et al. [7] proposed a Temporal Segment Network (TSN) for movement recognition in films, achieving state-of-the-art performance on the HMDB51 and UCF101 datasets. However, it requires significant computational resources and performs poorly with low-quality video inputs.

Lei Wang et al. [8] Proposed a light-weight structure for human action popularity the use of RGB data. Convolutional Neural Networks (CNNs) extract spatial features from video frames, even as ConvLSTM and FC-LSTM seize temporal movement functions. A Temporal-Smart Attention Model allows the network cognizance on important elements of the frames, improving accuracy and lowering noise. The Joint Optimization Module explores relationships among the outputs of the LSTM networks. Their technique finished sizable enhancements in movement reputation performance.

## III. PROPOSED SYSTEM

The proposed system for detecting human activity in infrared videos is designed to work well in low or no light conditions as shown in Fig. 1, the process from down by insertion of an infrared video image, followed by subtraction. CNN is used to extract spatial features focusing on patterns such as body position, while temporal features capture time-dependent features in motion in frame These features are provided to the ConvLSTM[11] model, which covers space and time analysis together to classify activities such as standing, sitting, kicking, walking of the desired data The system is trained and tested through use. Theoretically, the trained model processes unrecognizable video, provides activity labels, and generates annotations, indicating acceptable activity This system ensures accurate recognition and efficiency in use in monitoring and tracking.

### A. ARCHITECTURE

The Conv-LSTM[10] framework as shown in Fig. 2 effectively combines convolutional neural networks (CNN) and

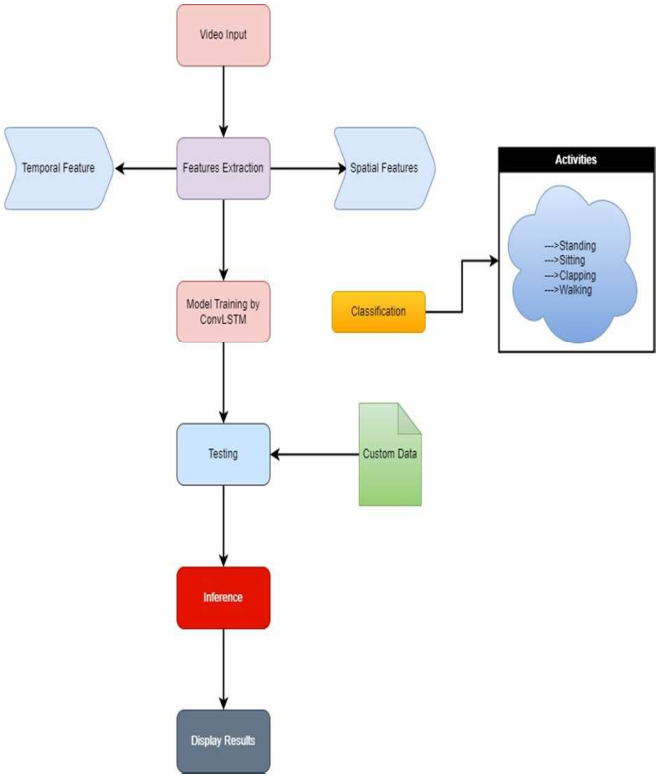



Fig. 1. Methodology

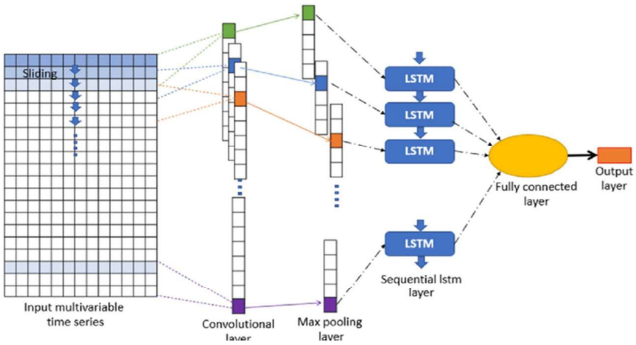



Fig. 2. Architecture of Conv-LSTM

long-term memory (LSTM)[11] networks to process multivariate time series data to identify human activity The process begins with a sliding window technique that partitions input time-series data into overlapping windows. These blocks then pass through convolutional layers, which extract spatial features, followed by maximum pooling layers that preserve important information and reduce spatial scale and then transfer the extracted spatial features to LSTM[11], which capture latency so are sequential patterns in the data. Finally, the fully connected layer combines features learned from convolutional and LSTM layers to produce a final output, which represents predicted activity This algorithm handles spatial and temporal aspects of human activity handle well, making it ideally suited for infrared video-based recognition tasks

### B. Data-Set Collection:

The dataset for this study was carefully constructed by capturing infrared (IR) videos from CCTV images. This method was chosen to overcome common challenges in traditional video-based activity detection, such as low light conditions, occlusions and by using IR cameras, the dataset captures temperature handwriting well in environmental conditions ensures clarity and accuracy

### C. Activities Recorded

The dataset includes seven distinct human activities, carefully selected to represent a range of common movements. These activities are:

1) Walking
2) Running
3) Sitting
4) Standing
5) Kicking
6) Punching
7) Jumping

### D. Data Collection Process:

The IR cameras were positioned in different ways, both indoors and outdoors, to simulate real-world monitoring conditions. A group of people with different physical characteristics, including age and gender, performed each task. The video was recorded at different locations, distances, and lighting conditions to ensure dataset robustness and scalability.

### E. Dataset Characteristics:

**Total Videos:** 15 video clips
**Duration:** Each video clip ranges between 20 to 25 seconds.
**Resolution:**The videos were recorded in standard CCTV resolution of 704x480 pixels, ensuring compatibility with typical surveillance systems.
**Annotation:**All activities were meticulously annotated by trained personnel to maintain high labeling accuracy.

This dataset provides a solid foundation for training and testing Human Activity Recognition (HAR) models, with a focus on surveillance applications utilizing IR imaging technology.

### F. Data Set PreProcessing

To improve the performance of the Conv-LSTM model for identifying multi-human activity in infrared videos, we applied basic data preprocessing steps. First, the images were resized to their default shape and normalized to [0, 1] for continuous training. Data enhancement techniques such as rotation, flipping, and brightness adjustment were used to improve normalization and reduce overfitting. In the posture calculation, 17 points representing body parts were extracted from each frame, organized into feature sets, labeled with activities such as standing, sitting, walking, etc. Then they were distributed the data set was divided into 85% for training and 15% for testing, providing robust model training and analysis. These preliminary steps refined the dataset, enabling the Conv-LSTM model[14] to better capture the spatial and temporal dependence for more accurate activity detection in real-world settings.

## IV. ALGORITHM

1: **Frame Extraction:** Extract frames $\{F_1, F_2, \ldots, F_n\}$ from $V$.
2: **Preprocessing:** For each frame $F_i$:
3: 1. Apply noise reduction.
4: 2. Normalize pixel values.
5: 3. Perform histogram equalization.
6: 4. Enhance contrast.
7: **Feature Extraction:** Use CNN to extract spatial features from preprocessed frames.
8: **Temporal Analysis:** Use LSTM to analyze temporal dependencies in the spatial features.
9: **Activity Classification:** Classify activities (e.g., walking, sitting) using a softmax layer.
10: **Output Generation:** Annotate frames with activity labels and compile into an output video.
11: **Performance Evaluation:** Calculate accuracy, precision, recall, and F1-score to evaluate the system.

## V. EVALUATION METRICS

Evaluating Human Activity Recognition (HAR) in infrared surveillance includes assessing the version's ability to as it should be located and classify human actions beneath various thermal conditions. Traditional assessment metrics including Precision, Recall, and F1 Score continue to be essential, however extra concerns like thermal image excellent, occlusion managing, and environmental adaptability play an important role in figuring out the effectiveness of HAR in infrared settings.

**Precision:**

measures how accurately the model identifies tremendous instances. It represents the share of correctly expected high-quality cases out of all times categorized as wonderful. A excessive precision value indicates that the version minimizes fake positives, making sure that most detected activities are definitely relevant. This is specially vital in surveillance programs, in which decreasing false alarms is important to keep away from needless safety responses. In an infrared-based totally HAR device, accomplishing excessive precision approach the model successfully distinguishes between actual human sports and heritage noise or non-human moves, thereby enhancing the reliability of the surveillance device.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

**Recall (Sensitivity):**

Recall measures the model's ability to identify all relevant instances in a dataset. It answers the question: "How many of the actual positives did the model correctly identify?" A high recall rate means the model is effective at detecting positive examples, which is critical in applications such as

fraud detection or disease diagnosis, where missing a positive instance can have serious consequences.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

**F1 Score:**

The F1 score provides a single metric that balances precision and recall, offering an overall measure of the model's predictive quality. It is particularly useful when there is a need to balance false positives and false negatives. Since it is the harmonic mean of precision and recall, the F1 score ensures that both metrics contribute equally to the overall evaluation. This makes it a reliable indicator of a model's effectiveness, particularly in cases involving imbalanced datasets where one class is significantly underrepresented.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Confusion Matrix**:

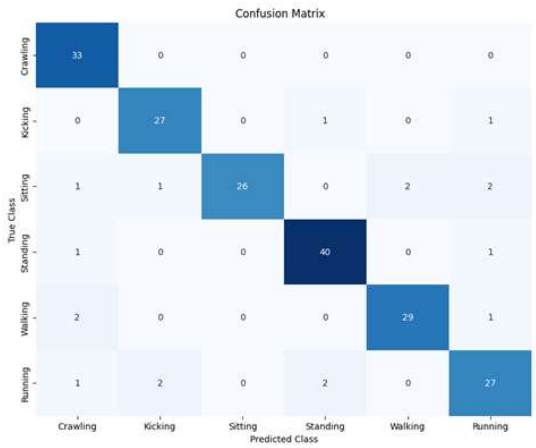



Fig. 3. Confusion matrix

Fig. 3 tells about the confusion matrix and provides a comprehensive view of the model's performance by comparing the predicted functions with the true functions Diagonal elements represent the best classified observations of each function, while diagonal elements indicates that it is well distributed. It is clear that the model exhibits high accuracy in identifying functions such as "standing" and "walking", with a significant number of correct predictions in the diagonal part but small incorrect classifications, such as "running" a it is confused with "walking" and sometimes the error in identifying "kicking".This misclassification may be due to the similarity in the positional dynamics between these functions. The uncertainty matrix, as shown in Fig. 3, provides valuable insight into the strengths of the model and areas for improvement, which helps to refine the usage recognition system for increased performance.

## VI. ROC CURVE:

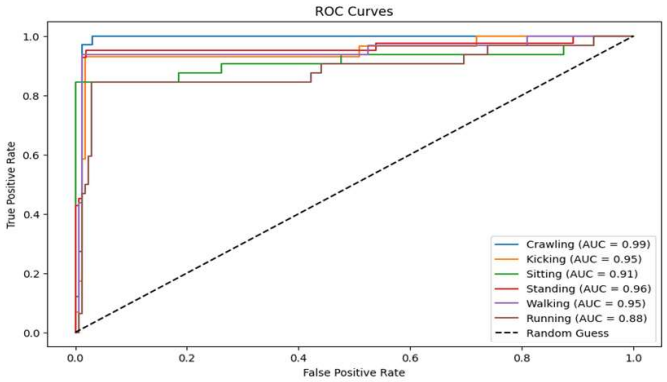



Fig. 4. ROC Curve

The receiver operating characteristic (ROC) curve as shown in Fig. 4 provides a graphical representation of the model's ability to discriminate between activity classes by plotting the false positive rate (FPR) against the true positive rate (TPR) at various thresholds. Functions such as "crawling" and "standing" exhibit excellent AUC values of 0.99 and 0.96, respectively, indicating high classification accuracy. Other activities such as "kicking" and "walking" also perform well with an AUC value of 0.95, while "Sitting" and "Running" show slightly lower AUC values of 0.91 and 0.88, respectively, indicating potential areas for improvement The dashed diagonal line represents the efficiency of random classification, which serves as the starting point. The ROC curve, as shown in Fig. 4, reveals the strong performance of the model in most applications, and indicates which applications may need further fine-tuning to increase overall system efficiency.

## VII. RESULTS AND ANALYSIS

The performance of the Conv-LSTM model has evaluated on a take a look at set inclusive of infrared films with annotated key points. Despite the demanding situations posed by using low-light situations, the model confirmed sturdy performance in as it should be detecting and localizing the 17 key points at the human frame. The infrared imagery, even though tormented by reduced lighting fixtures, enabled the model to preserve unique spatial mapping, which is essential for the subsequent interest classification. This sturdy performance beneath hard conditions highlights the version's effectiveness in actual-international eventualities, wherein lighting fixtures can be suboptimal.

### *A. Accuracy and precision curve:*

Fig. 5 Tells about the accuracy and precision curves that offer a complete evaluation of the model's overall performance throughout various selection thresholds. The accuracy curve (depicted in blue) represents the share of accurate predictions made with the aid of the model out of all predictions,

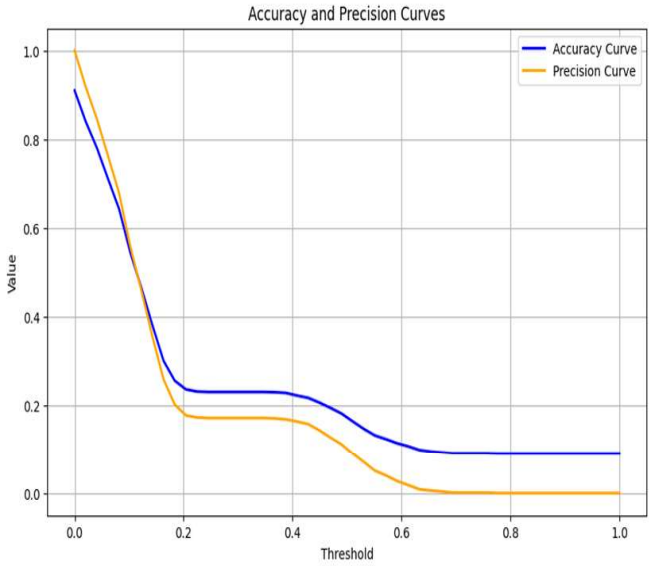



Fig. 5. Accuracy and Precision Curves

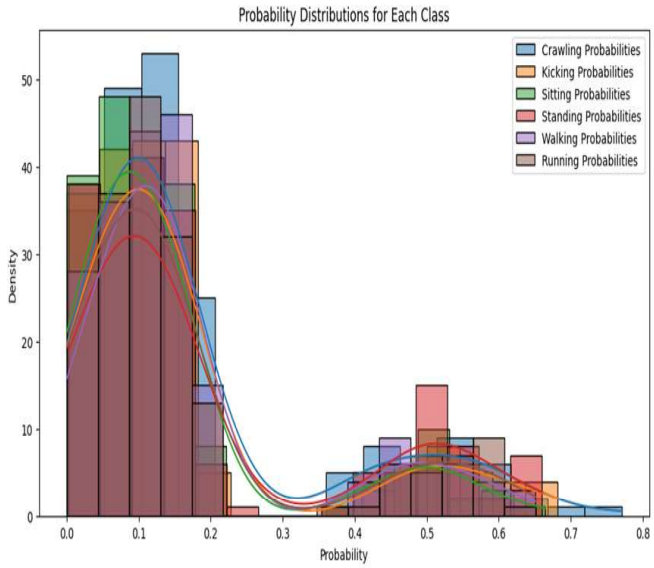



Fig. 6. Probability Distribution Curves for Each Class

highlighting its general reliability in class duties. On the other hand, the precision curve (depicted in orange) measures the share of actual fantastic predictions among all positive predictions, imparting insight into the version's capability to keep away from false positives.

As the selection threshold increases, both accuracy and precision reveal an exceptional decline. This conduct displays the intrinsic trade-off between false positives and false negatives, a commonplace assignment in classification tasks. At decreased thresholds, the version achieves high accuracy and precision, indicating its ability to make effective predictions with minimal errors. However, as the threshold fee increases, the stringent classification criteria cause a lower performance, with a growing probability of misclassifications.

The curves function as a crucial tool for identifying the optimum threshold value that balances accuracy and precision. This stability ensures that the class system maintains robustness and reliability, in particular within the context of the interest recognition device being evaluated. By studying those curves, builders, and researchers can high-quality-track the model to reap the desired overall performance metrics, addressing specific use-case requirements.

Fig. 5 correctly illustrates this evaluation, providing a visual illustration of the interplay among accuracy and precision because the choice threshold varies. This enables stakeholders to make informed choices approximately the brink of putting, in the end enhancing the type machine's universal efficacy.

### *B. Probability Distribution*

The possibility distribution plot in Fig. 6 tells about the likelihoods for numerous human activities like Sitting, Standing, Running and Kicking in infrared videos and processed using a Conv-LSTM version. Each colored curve represents the possibility density for the respective activities, whilst the overlaid histograms offer insights into the underlying data distribution. The primary objective of this analysis is to assess the predictive performance of the Conv-LSTM version in recognizing human activities from infrared video records. Key observations reveal that activities together with standing and walking showcase awesome distribution peaks, indicating steady predictions. However, overlaps among similar sports, such as status and strolling, suggest ability demanding situations in differentiating between them. Additionally, the presence of outliers, with decrease or better chances for unique sports, highlights occasional misclassifications or uncertain predictions. This analysis underscores the significance of addressing overlaps and outliers to in addition beautify the version's reliability and robustness in activity reputation duties.

### *C. Comparative analysis*

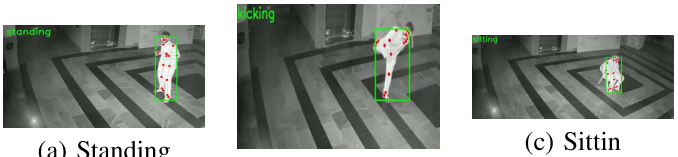
TABLE I
PERFORMANCE COMPARISON OF VARIOUS MODELS

| **Model** | **Accuracy (%)** | **Dataset Used** |
|---|---|---|
| Conv-LSTM (Our Study) | 97.71 | Infrared videos collected dataset |
| **[1]** Enhanced YOLOv5 | 95.5 | Diverse surveillance datasets |
| **[4]** Attention-based LSTM | 93.0 | Abnormal activity datasets |
| **[9]** Deep Conv-LSTM | 92.0 | Infrared images |
| **[10]** Action Recognition from Thermal Videos | 90.5 | Thermal video datasets |

Table I offers a performance assessment of various hobby reputation models on special datasets. The Conv-LSTM version from this examination carried out the highest accuracy of 97.71% on an infrared video dataset, outperforming Enhanced

YOLOv5, Attention-based LSTM, Deep Conv-LSTM, and Action Recognition models based totally on thermal videos. This highlights the effectiveness of the Conv-LSTM method in infrared-primarily based pastime reputation obligations.

### *D. Outputs and Results*




(a) Standing (b) Kicking (c) Sittin

Fig. 7. Actions detected in different scenarios.

The diagrams in Fig. 7 tell about how human actions are detected and classified using computer vision modeling. The first image 7(a) shows a person performing the actions, where the prototype searches for points on the body carefully and covers them, the second image 7(b) tells about focuses on the dynamic motion trees by skeleton-based definition and bounding box. Standing refers to a person's posture, including the model's ability to accurately identify and classify standing positions The 7(c) tells about in the exhibit shows a person sitting sitting conditional, highlighting the versatility of modeling in the recognition of body structures. These examples highlight the robustness of the model in identifying and distinguishing the range of human behavior in monitoring environments, leading to reliable results in real-world situations

## CONCLUSIONS AND FUTURE WORK

The challenge efficaciously carried out a real-time human hobby reputation (HAR) gadget for infrared (IR) video analysis, utilising present day laptop imaginative and prescient and deep mastering techniques. By leveraging Convolutional Long Short-Term Memory (Conv-LSTM) networks, the machine excels in extracting and classifying spatiotemporal functions, enabling the accurate identity of diverse human activities, which include crawling, kicking, sitting, standing, taking walks, and going for walks. The capability to seize both spatial and temporal data inside the infrared spectrum enhances the model's performance in dynamic and challenging environments, allowing it to provide real-time pastime detection.

Despite the achievement of the cutting-edge machine, its capability is currently confined by the dataset used for schooling, which most effectively includes a restricted set of six activities. This dilemma is because of the reliance at the nice, range, and amount of the dataset, which restricts the version's generalizability to a much broader range of human moves. Therefore, one of the key regions for development is to increase the dataset, incorporating a broader spectrum of human activities and postures, ensuring the version is capable of recognizing a whole lot of behaviors in actual-world scenarios.

In addition, the system's future iterations will intention to combine extra advanced techniques, along with pose estimation, a good way to allow for a greater designated information of the concern's body posture and moves. Incorporating anomaly detection algorithms to perceive uncommon or suspicious activities will even enhance the system's versatility in regions like safety and surveillance, where detecting unusual behaviors is important.

Furthermore, improving type accuracy will beautify the version's ability to differentiate among diffused sports or movements that could appear similar, in the long run minimizing fake positives and enhancing the system's reliability. By addressing these modern limitations, the gadget has the ability to turn out to be more robust and adaptable to a huge range of use instances, inclusive of public safety tracking, human behavior analysis, and advanced protection systems.

## REFERENCES


[1] U. Gawande, K. Hajari, and Y. Golhar, "Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map," *Artif. Intell. Rev.*, vol. 57, no. 2, p. 16, 2024.

[2] A. K. Jhapate, S. Malviya, and M. Jhapate, "Unusual crowd activity detection using OpenCV and Motion Influence Map," in *Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA)*, Feb. 2020, pp. 1–6. (pp. 1-6). IEEE.

[3] J. H. Jeong, H. H. Jung, Y. H. Choi, S. H. Park, and M. S. Kim, "Intelligent complementary multi-modal fusion for anomaly surveillance and security system," *Sensors*, vol. 23, no. 22, p. 9214, 2023.

[4] M. Kumar, A. K. Patel, M. Biswas, and S. Shitharth, "Attention-based bidirectional-long short-term memory for abnormal human activity detection," *Sci. Rep.*, vol. 13, no. 1, p. 14442, 2023.

[5] A. Mehmood, "Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks," *IEEE Access*, vol. 9, pp. 138283–138295, 2021.

[6] N. Jaouedi, N. Boujnah and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447-453, 2020.

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," *European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016, pp. 20-36, doi: 10.1007/978-3-319-46484-8$_2$.

[8] L. Wang et al., "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE Access*, vol. 6, pp. 17913-17922, 2018, doi: 10.1109/ACCESS.2018.2811486.

[9] A. Akula, A. K. Shah and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cognitive Systems Research*, vol. 50, pp. 146-154, 2018.

[10] A. Sharma and R. Singh, "ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 34, 2023.

[11] M. Kumar, B. Murugan and S. Pooja, "Enhancing human activity recognition through deep learning: Comparative analysis of single frame CNN and convolutional LSTM models," *2024 9th International Conference on Control and Robotics Engineering (ICCRE)*, Osaka, Japan, 2024, pp. 400-405, doi: 10.1109/ICCRE57820.2024.10546789.