scientific reports

Check for updates

OPEN

Bi-directional ConvLSTM networks for early recognition of human activities and action prediction

M. Ashwin Shenoy[1,2]✉, N. Thillaiarasu[1], S. Santhosh[3] & S. Sandeep Kumar[4]

Early detection of human activity is essential in domains including robotics, entertainment, surveillance, and healthcare. Early detection that is accurate enables prompt decision-making, enhancing system responsiveness and overall effectiveness. Conventional action recognition techniques can't handle sequential and incomplete data well since they are usually built for offline analysis and concentrate on detecting entire actions. Early detection necessitates real-time result prediction and incomplete activity identification, which are difficult for many current models to do. In order to enhance early detection and prediction of human behaviors, this study proposes a unique method utilizing a Bi-Directional Convolutional Long Short-Term Memory (Bi-ConvLSTM) network. By incorporating both spatial and temporal connections, the model processes sequential data and makes it possible to identify activity initiation and continuing activities with greater accuracy. By examining the temporal sequence of input frames, the Bi-ConvLSTM network is intended to identify the beginning of an activity and forecast its course. The proposed approach utilizes a segment-based strategy in which the input sequence is broken down into smaller intervals, allowing the model to focus on specific temporal segments. This improves the network's capacity to recognize tiny motion patterns and contextual signals indicating the start of an activity. The model is tested on a real-world dataset that includes a variety of human behaviors recorded in complicated contexts. Experimental findings show that the proposed Bi-ConvLSTM model outperforms current models such as CNN, InceptionV3, VGG19, and regular ConvLSTM networks, with an average accuracy of 89.54%. The findings show that the Bi-ConvLSTM model efficiently balances early detection accuracy with decision-making speed, making it appropriate for real-time applications. This study demonstrates the ability of Bi-ConvLSTM networks to improve early detection and prediction of human behaviors, opening the door for more responsive and intelligent systems.

Keywords Early Recognition, Sequential Data Analysis, Bi-directional ConvLSTM Network, Activity Onset Estimation, Segment-Based Models

Numerous applications in a variety of sectors, such as robotics, entertainment, surveillance, and healthcare, among others, demand early detection. There are several methods for detecting human behavior, most of them concentrate on enhancing offline analysis precision. Since current action recognition algorithms frequently learn to recognize entire actions only after the acts have finished and all necessary information has been gathered, they are restricted in their ability to handle sequential data efficiently. Recognizing incomplete activities and the categories that go along with them is essential for early detection. Unfortunately, during their training phase, many popular action recognition algorithms sometimes fail to recognize incomplete actions. Unlike earlier processes, early identification prioritizes speed over accuracy when making decisions.

The difference between activity categorization and activity prediction is seen in Fig. 1. Activity categorization ensures high accuracy at the expense of delayed decision making by identifying an action after it has been finished. Activity prediction, on the other hand, seeks to detect an action in progress before it is finished, allowing for real-time decision-making at the possible expense of accuracy. This essential distinction emphasizes the necessity of early detection strategies that can successfully manage insufficient data.

[1]School of Computing and Information Technology, REVA University, Bangalore, Karnataka, India. [2]Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Department of CSE, Nitte, Karnataka, India. [3]Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Department of ISE, Nitte, Karnataka, India. [4]Computer and Communication Engineering, Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Department of CCE, Nitte, Karnataka, India. ✉email: ashwinshenoy14@gmail.com; ashwin.shenoy@nitte.edu.in

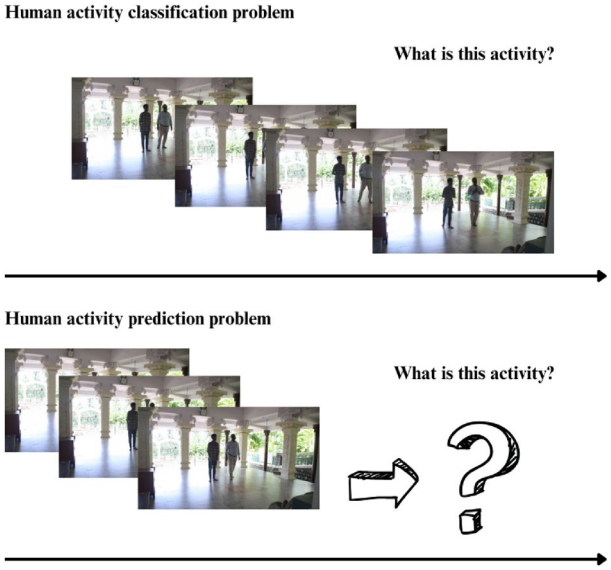


Fig. 1. The distinction between Activity Classification and Activity Prediction Challenges.

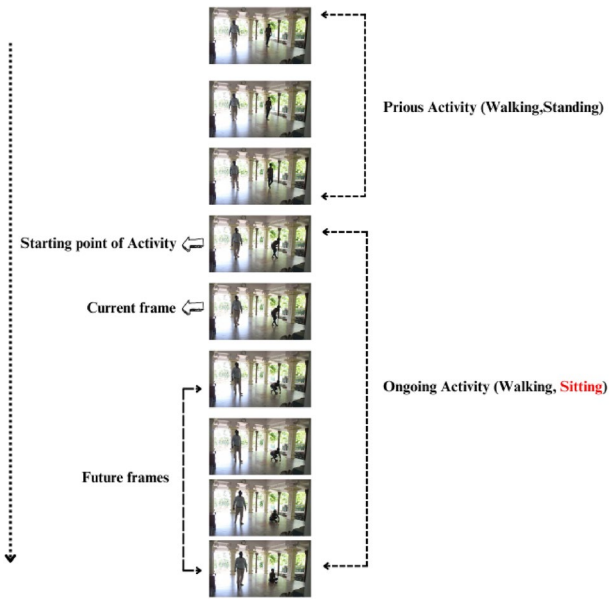


Fig. 2. Overview of the working of the proposed system.

It represents a burgeoning area, with some recent contributions[1–8] since its inception. The distinction between anticipating body motion and categorizing the type of body motion is critical. Unlike the latter, which studies both current and historical data, the former predicts future occurrences. Recognizing the many kinds of human activity is an important step. As discussed in [4,5, 8, 7, 9], several strategies for early detection of human activities depend on detecting the action category of body movements before the activity is completed. We explore several famous action recognition approaches and evaluate their ability to recognize and categorize current actions. Several techniques [10–13] established for early detection assume that the activity has begun. This assumption directs computational training.

Examine the several significant activity identification approaches and examine their ability to recognize and categorize ongoing actions. Several approaches [10–13] aimed for early detection assume that the present action has already started. This assumption promotes the development of computer models designed to properly categorize incomplete actions. This study focuses on the early detection of activities at the commencement phases of actions, using segment-based models[3]. The research presents a substantial breakthrough in a strategy for predicting the start of continuing actions, as seen in Fig. 2. The approach employs a bi-directional network

based on Convolutional Long Short-Term Memory(ConvLSTM) to create an accurate prediction about the start of the action sequence, with a margin of error ranging from five to ten frames.

This paper addresses these limitations by proposing a novel Bi-Directional Convolutional Long Short-Term Memory (Bi-ConvLSTM) model for early action detection and prediction. Unlike traditional ConvLSTM or convolutional neural network (CNN) based models, the proposed approach leverages temporal dependencies in both forward and backward directions, enhancing its ability to detect activity onset within a short time window of 5–10 frames. Our approach also adopts a segment-based strategy, allowing the model to focus on fine-grained temporal slices to identify subtle motion patterns signaling the start of an activity.

The remainder of this paper is organized as follows: Chapter 2 presents the background and related work, providing a foundation for understanding the key concepts and existing approaches in the domain. Chapter 3 details the proposed methodology, describing the framework, algorithms, and techniques used in this study. Chapter 4 discusses the experimental results, analysis, and key observations derived from the implementation of the proposed system. Finally, Chapter 5 concludes the paper by summarizing the main findings and highlighting potential directions for future research and improvement.

Background

Several computational models created to detect human activity may be found in the computer vision literature. As noted in [17, 4–6, 8–9, 23, 26, 28–29], recent initiatives have attempted to solve the difficulty of early identification, even though many existing models favor accuracy in offline processing over decision-making speed. Notably, though, none of these studies–including one earlier attempt[5]–have looked into the possible advantages of movement prediction. Prior studies have primarily focused on evaluating the advantages of different feature encodings and classifiers for early identification. Such research was carried out by Ryoo[6], who proposed integrated and dynamic bag-of-word models to help identify human interaction early. Furthermore[17], trained the classifier for enhanced recognition using a unique loss function. Despite these advantages, the potential advantages of integrating movement prediction into early detection systems were not taken into account by any of the previously proposed alternatives. A latent variable method was developed by Wang and associates[20] to infer unknown human behaviors.

The way these methods depict and identify partial actions differs from mine. These methods seek to forecast a human subject's course or destination, especially to predict actions or activities that will be derived from the subject's trajectory or goal. This chapter goes beyond merely classifying and forecasting planar or three-dimensional trajectories. Although the benefits of early identification have not been examined in prior research, studying human motion dynamics is an important field of study. Wang and associates[20] examined nonlinear time series using Gaussian process dynamical models. Using a low-dimensional latent space with dynamics, Wang and associates mapped the latent space to an observation space. In another study, Cao and Nevatia[21] inferred postures and motions using force analysis. The motion prediction method integrated accelerations and forces to compute the 3D positions of the posture in each frame. However, it should be noted that long-term forecasting dependencies were not taken into consideration by this approach. Pavlovic and his colleagues[22] were able to create models of human behavior by switching linear dynamic system models. They employed these techniques for offline data segmentation and identification rather than motion prediction and early pattern recognition.

Early human activity recognition

The goal of early recognition in computer vision is to quickly recognize an action by examining partial action sequences. Several models for human action recognition have been developed via computer vision. The study offers a novel technique for real-time action prediction with a low-cost depth camera. It gets over this restriction by using soft label learning for subsequences, which is different from current systems in that it doesn't presuppose knowledge of the progress level. When compared to other models, the suggested regression-based model that uses the effective local accumulative frame feature (LAFF) performs better on RGB-D sequences, demonstrating its usefulness in practical applications.

The paper investigates the challenging challenge of live action recognition from streaming 3D skeletal data using a novel multi-task Joint Classification-Regression Recurrent Neural Network. The model[11] automatically captures complicated long-range temporal dynamics and uses deep LSTM subnetworks to identify time properly. Traditional sliding window techniques are no longer necessary because of the shared optimization target, which increases efficiency. The effectiveness of the suggested approach is shown through experiments on both the Gaming 3D Dataset(G3D) dataset and a new dataset.

A key technique for improving a robot's ability to recognize human activity during encounters is presented in the paper[12]. The method used in first-person movies emphasizes early awareness by succinctly expressing observations prior to the start of an action by utilizing the idea of "onset." By integrating event history and visual data, the technique helps the robot to anticipate and respond to typical human actions more quickly. The outcomes of the tests show how this method enhances and expedites identification. The difficult challenge of identifying human behaviors in partially viewed movies, which have real-world implications, is addressed in the work[13]. By segmenting activities, employing spatiotemporal characteristics with sparse coding, and combining the likelihoods of the segments to determine a global posterior for the activities, the proposed method decomposes the issue into a collection of probabilities. Activities with notable intra-class variations can be better represented by an extension that combines segments of different temporal lengths. The review of real films has demonstrated the effectiveness of the suggested methods. These suggestions have shaped current state-of-the-art methods in a number of contexts, such as completely observed films and activity prediction.

Wang and associates[23] examined human behavior in actual movies, paying special attention to instances when inconsequential elements were prevalent in human behavior. They trained their algorithm on the Action Thread dataset, which necessitated classifying every shot separately. The benefits of eliminating non-action

scenes from videos, particularly those that don't feature human motion, were examined by the writers. In order to address this problem, they included a non-action classifier designed to lessen the significance of unimportant video clips. The overall performance of action detection systems was enhanced by the classifier's capacity to consistently identify frames devoid of any activity.

The authors used LSSVM, or least-squares support vector machines, to achieve recognition. The paper[24] used switching linear dynamic system models to study the learning models associated with human dynamics. Their research shows that these models are useful for assessing figure motion, highlighting their significance for tasks such as summarizing approximation inference techniques and identifying gestures. A variational inference approach was also described in the study, emphasizing the benefits of seeing conventional statistical models as mixed-state graphical models. Even if these methods are used for offline recognition and segmentation, it's important to be aware of their limitations in early recognition applications.

Despite having a relatively low observation ratio in the video sequence[25], novel action anticipation method showed impressive prediction accuracy. In many stages, the researchers created an advanced LSTM framework with capabilities that comprehend the surroundings and the ongoing behaviors. They also presented a novel loss function designed to encourage the algorithm to predict the correct class as quickly as possible. In terms of early prediction, the system achieved an improvement in accuracy of 22.0% on Joint-annotated Human Motion DataBase 21 action categories(JHMDB-21), 14.0% on University of Texas(UT)-Interaction, and 49.9% on UCF-101, surpassing the most advanced action prediction techniques.

An architectural framework that makes use of knowledge distillation was presented by[27]. The early detection-designed network serves as the foundation for the student prototype in this framework. The required guidance for learning is given by a skilled instructor model that foresees future events and incorporates additional information about the task being studied. This method leverages the benefits of semi-supervised learning by using both labeled and unlabeled training data. In the evaluation of the Nanyang Technological University(NTU) RGB-D dataset, our solution outperformed the LSTM and Recurrent Neural Network(RNN) approaches, achieving an Area Under the Curve(AUC) of 62.8% on the Receiver Operating Characteristic Curve(ROC) curve. An action recognition network oversees the training of an action anticipation network in a novel knowledge distillation method introduced by[28]. In order to accurately predict future occurrences, this instruction helps the anticipation network focus on important information. Using unlabeled data and a self-supervised learning approach, create the loss function to manage changes in semantic ideas in movies.

Performance improved significantly, reaching 75.8%, once the loss function was replaced. Accuracy significantly improved with the addition of a symmetric bidirectional attention loss, surpassing the previous best result with 76.6% on the JHMDB dataset. We attributed the success to the combination of optical flow and RGB data. Wang and colleagues[29] emphasized the significance of pinpointing the initiation of an action.

To determine the likelihood that a certain frame would serve as the starting point, they created a bidirectional RNN technique. By analyzing the dynamics of the acts that before and followed the frame, this was achieved. Their method, which employed a bidirectional LSTM, effectively preserved two separate information flows: forward progression and backward traverse. By obtaining an AUC of 61.2%, they demonstrated the efficacy of their approach on the Montalbano Gesture dataset and showed its advantage in situations with unclear beginning positions.

Human Activity Recognition (HAR) using deep learning[30], particularly with CNNs and LSTMs, has improved the recognition of complex tasks. Traditional sensor-based systems often misclassify intricate activities due to sensor inaccuracies. Vision-based systems, leveraging deep learning, enhance accuracy and cost-efficiency by analyzing visual data, reducing reliance on faulty sensor readings and improving overall performance in recognizing complex activities. The paper explores human activity recognition (HAR) using deep learning on image data[31], employing transfer learning and ensemble techniques with models like Visual Geometry Group 16-layer network(VGG16), Residual Network with 50 layers(RESNET50), and Efficient Neural Network, variant B6(EfficientNetB6) to improve accuracy and efficiency. Deep learning enhances HAR by identifying patterns in image data, though variations in human shape and motion pose challenges. Training these models also demands significant computational resources. The paper explores human activity recognition (HAR) using deep learning, particularly CNNs and RNNs[32], to analyze wearable sensor data and capture spatial and temporal dependencies in activities like walking, running, and sitting. It discusses the accuracy of deep learning models, challenges in computational complexity and scalability, and highlights real-time performance as a key area for future research.

The paper examines Human Activity Recognition (HAR) using LSTM and Gated Recurrent Unit(GRU) models[33], achieving 80–85% accuracy but underperforming compared to logistic regression and SVM models, which reached 93–94% accuracy. Challenges include handling diverse activities and limited data samples, affecting deep learning model performance and early activity recognition. The paper explores human activity recognition (HAR) in smart homes using the Fully Convolutional Network with Long Short-Term Memory(FCN-LSTM) model[34], which effectively captures spatial and temporal variability in activities like walking, sitting, and cooking, enhancing recognition accuracy through deep learning. The study explores deep learning-based human activity recognition using accelerometer and gyroscope data to identify six activities. A 1D-CNN-BiLSTM model[35] demonstrated high accuracy, especially for walking-related actions, despite limitations from a small dataset and low activity variability.

This comprehensive study looks at many computer models and methods in the subject of human action detection, with a focus on early identification. Existing techniques usually prioritize offline processing accuracy, even if recent efforts emphasize the need of movement prediction for better early detection. New techniques are presented in the reviewed literature, such as the use of an online action detection system that combines a low-cost depth camera for real-time prediction with a Joint Classification-Regression Recurrent Neural Network algorithm. Advances in action anticipation, trajectory analysis, and information distillation all of which exhibit

Paper	Key Points	Methodology	Datasets	Results
11	Early recognition of human actions using a depth camera, no progress level assumption, soft label learning for subsequences	Regression-based model with Local Accumulative Frame Feature (LAFF) and Joint Classification-Regression Recurrent Neural Network with deep LSTM subnetworks	New dataset, G3D dataset	Outperformed existing models on RGB-D sequences
12	Enhancing robot recognition of human activities using first-person films, early recognition via the 'onset' concept	Combines event history and visual data	Not specified	Improved and sped up recognition
13	Recognizing human activities in partially observed videos	Segmentation of activities into spatiotemporal features with sparse coding, global posterior for activities	Actual videos	Successful evaluation in activity prediction and fully observed videos
23	Human behavior recognition in real films, removal of non-action parts	Non-action classifier to reduce the importance of irrelevant segments, LSSVM	Action Thread dataset	Improved action detection performance
24	Learning models for human dynamics using switching linear dynamic system models	Variational inference method for mixed-state graphical models	Not specified	Effective in analyzing figure motion and gesture identification
25	Action anticipation with a low observation ratio	Sophisticated LSTM framework, innovative loss function	JHMDB-21, UT-Interaction, UCF-101	Accuracy improvement of 22.0%, 14.0%, and 49.9% respectively
27	Architectural framework with knowledge distillation for early detection	Semi-supervised learning, teacher-student model	NTU RGB-D dataset	AUC of 62.8%, outperformed LSTM and RNN methods
28	Knowledge distillation for action anticipation network training	Self-supervised learning, symmetric bidirectional attention loss	JHMDB dataset	Accuracy of 76.6%, surpassing the previous best result
29	Pinpointing initiation of action using bidirectional RNN	Bidirectional LSTM for forward and backward information flow	Montalbano Gesture dataset	AUC of 61.2%, superior in ambiguous starting points

Table 1. Summary of Early Activity Detection Methods.

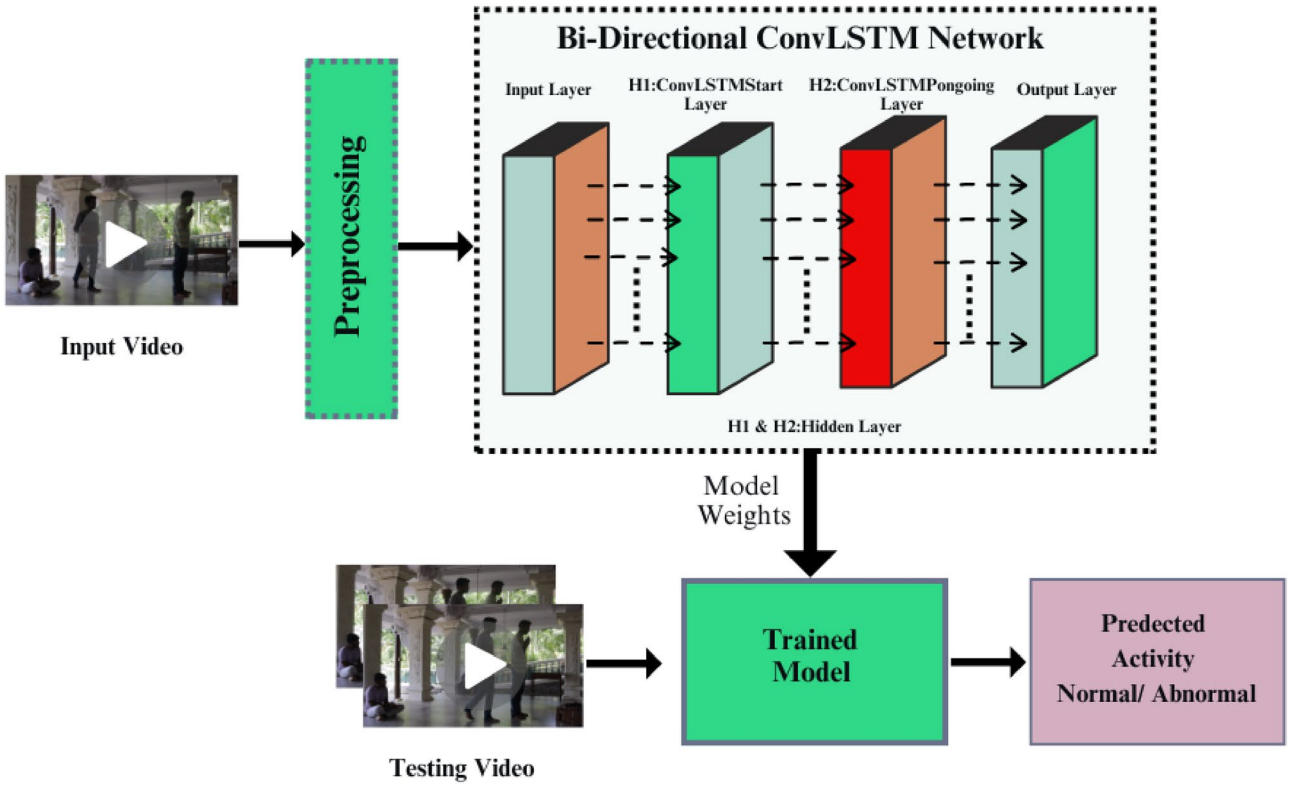


Fig. 3. Proposed Bi-ConvLSTM architecture for Predicting Ongoing Activity.

improved accuracy and speed of decision making are also influencing the evolving area of early identification. An overview of early activity detection techniques is shown in Table 1.

Proposed methodology

The proposed Bi-ConvLSTM architecture for anticipating ongoing activity is shown in Fig. 3. The model is made up of several layers.

Dataset

The temple dataset[2] was chosen for this study due to the unique and complex nature of human activities occurring within a temple environment. Temples are dynamic spaces where both normal and unusual human behaviors can be observed, making them an ideal setting for studying human activity recognition. Unlike other public or controlled environments, temples involve a wide range of cultural and social interactions, including walking, standing, sitting, praying, and abnormal activities such as pickpocketing or chain snatching. The crowded and diverse nature of temple gatherings introduces significant variability in terms of illumination, background complexity, and human movement patterns. This diversity reflects real-world challenges associated with human activity recognition, such as occlusion, cluttered backgrounds, and varying perspectives. The dataset, therefore, is well-suited for developing robust deep learning models capable of accurately recognizing and classifying human activities in complex, real-life scenarios. Additionally, the combination of both structured and spontaneous behaviors enhances the dataset's ability to capture natural variations in human actions, improving the model's capacity to generalize to unseen data.

The dataset used for this study was specifically collected from a temple environment to capture a wide range of human activities, both natural and staged. Data was collected from single and multiple subjects performing various actions within a single visual frame. A total of 50 video clips were recorded, each lasting between one to three minutes. These clips featured both pre-planned (natural) actions and spontaneous behaviors occurring in the temple setting. The collected data included a diverse range of activities such as standing, sitting, walking, praying hands, forward fall, fighting, chain snatching, and pickpocketing. To enhance the dataset's representativeness, data collection was conducted when over 50 individuals were present for more than five hours. This ensured the inclusion of diverse lighting conditions, crowd densities, and variations in human movement patterns.

The raw video data was processed using the FFmpeg tool to extract frames at a rate of three frames per second (FPS), reducing redundancy while capturing meaningful visual information. This extraction resulted in a total of 2,300 annotated frames representing both frontal and posterior views of human actions. After processing and frame extraction, the complete dataset comprised over 70,000 individual frames. These frames were annotated using the LabelImg tool, where each frame was assigned activity labels based on detected hand gestures and body postures. The annotations were stored in YOLO-compatible format (.txt files), with each entry containing the class name and the bounding box coordinates for object localization and activity classification.

To standardize the data, all extracted frames were resized to 224×224 pixels to fit the input size requirements of deep learning models. Data augmentation techniques were applied to increase the dataset's variability and improve model generalization. The augmentations included rotation (0°, 90°, 180°, 270°), flipping (horizontal and vertical), and brightness adjustment (between 0.7 and 1.3). Additionally, noise reduction techniques such as Gaussian blur and median filtering were used to improve image clarity by removing motion blur and background noise.

The criteria for removing duplicate or redundant frames from the dataset were based on a similarity threshold between consecutive frames. After extracting frames at a rate of three frames per second (FPS), a structural similarity index was computed between each pair of consecutive frames. If the similarity between two frames exceeded 80% (0.8), the latter frame was identified as redundant and removed. This approach ensured that only meaningful variations in activity were retained while reducing data redundancy. Lowering the frame rate to two frames per second helped further minimize duplication, improving the overall quality and diversity of the dataset for better model training and generalization.

The dataset was carefully organized into training and testing sets following an 80-20 split. The training set included 1,000 frames per activity, amounting to a total of 8,000 frames for all eight activities. The testing set consisted of 200 frames per activity, totaling 1,600 frames for model evaluation. This balanced split ensured that the model was trained on a sufficient volume of data while being evaluated on a diverse set of unseen samples. The dataset's variability was further increased by including images with occlusions (multiple individuals within a single frame), background clutter, and varying lighting conditions to simulate real-world challenges. This comprehensive and diverse dataset effectively supports the development of deep learning models for human activity recognition in temple environments.

Pre-processing

To ensure the data is in an appropriate format for efficient neural network processing, a pre-processing step is performed to prepare raw video frames for input into the ConvLSTM networks. This step consists of two main phases: frame extraction and frame resizing.

Frame extraction involves decomposing continuous video streams into individual still frames at a fixed frame rate (e.g., frames per second). This allows the neural network to analyze each frame independently and learn the temporal dynamics across sequential frames.

Frame resizing ensures that all extracted frames are scaled to a uniform resolution of 128×128 pixels, which facilitates consistent input dimensions and compatibility with the network architecture.

By performing these pre-processing operations, the raw video data is transformed into a standardized and structured format. This helps ensure that the subsequent ConvLSTM stages can operate effectively, contributing to more accurate detection and prediction of human activities within the video sequences.

ConvLSTM network for action starting point detection

The input sequence, represented as X = x1, x2,..., xT, is divided into discrete segments or intervals using the segment-based approach. A ConvLSTM network created especially for action beginning point identification is used to analyze each segment separately. A targeted analysis is made possible by this segmentation, in which the model produces a probability distribution, Pstart(t), for the possibility that an activity will begin at each time step t in the segment. The action beginning point detection can be expressed mathematically as follows:

$$P_{\text{start}}(t) = \text{ConvLSTM}_{\text{start}}(X_t) \tag{1}$$

where Xt represents the input at time step t.

The ConvLSTMstart network is ideally suited for identifying the start of an action within a sequence because it uses the architectural advantages of Convolutional Long Short-Term Memory (ConvLSTM) networks to handle spatial and temporal input in an integrated way. By adding convolutional operations to the input-to-state and state-to-state transitions, convLSTM layers expand on conventional LSTM and enable the model to capture temporal dependencies and maintain spatial linkages. With video data, where frames display temporal progression and spatial coherence, this design works very well.

The ConvLSTMstart model uses a segmentation approach to break up the input sequence into smaller, time-based segments in order to increase its precision. Because each segment is processed independently, the network may concentrate on certain temporal periods when action initiation is most likely to take place. By limiting the area of analysis, this method reduces processing cost while maintaining the capacity to detect minute motion changes, posture adjustments, or contextual modifications that signal the start of an activity. An essential part of the ConvLSTMstart network's architecture, the segmentation procedure maximizes the examination of spatial-temporal patterns. The network guarantees that each temporal slice is examined independently of irrelevant sequence parts by splitting the input into distinct segments. This avoids the potential for information dilution in a global study, where overlapping or continuing actions might mask important signals for initiating action.

Convolutional filters are used by the ConvLSTMstart network to analyze each segment, extracting spatial characteristics that are then temporally encoded by LSTM units. This pipeline preserves the integrity of spatial features while guaranteeing that temporal transitions within each segment are adequately preserved. The model is a highly specialized solution for tasks like surveillance, where precisely detecting the onset of abnormal behaviors is crucial, or human-computer interaction, where precise action timing enhances responsiveness. This is made possible by the segmentation-driven analysis, which also makes the model robust against noise or irrelevant activities and allows it to adapt to variable-length sequences.

ConvLSTM network for prediction of ongoing activity

To include just the frames from the starting point to the current time step, truncate the frame sequence based on the starting point identified by the ConvLSTMstart layer. The shortened sequence is sent to this network and is represented as:

$$X_{\text{truncated}} = x_{\text{start}}, x_{\text{start}+1}, \ldots, x_T \tag{2}$$

In this case, the anticipated beginning position is xstart(t). A probability distribution Pongoing(t) for the ongoing action is produced by the ConvLSTM network for prediction at each time step t:

$$P_{\text{ongoing}}(t) = \text{ConvLSTM}_{\text{ongoing}}(X_{\text{truncated}}) \tag{3}$$

where ConvLSTMongoing denotes the ConvLSTM network dedicated to ongoing activity prediction.

To address the challenge of early action recognition, the proposed Bi-ConvLSTM model is explicitly trained and evaluated using only the early segments of the video sequences rather than the entire duration. Specifically, the input video is segmented into temporal windows, and the model is trained to detect the starting point of an action using partial observations–typically only 20% to 40% of the full video sequence. Once the starting point is detected by the ConvLSTMstart network, the ConvLSTMongoing network processes only the truncated sequence from that point onward. This design simulates real-time, online scenarios where full video context is not available at prediction time. The results in the evaluation section are also presented based on performance using these early portions of video data, demonstrating the model's ability to accurately anticipate and classify activities at an early stage. This clearly distinguishes our approach from conventional full-sequence recognition models that rely on observing the complete action before making predictions.

Require: Sequence of frames $X = \{x_1, x_2, \ldots, x_T\}$
Ensure: Continuous prediction of activity
procedure PREDICTION(Activity)
  Divide the input sequence into smaller segments
  for each segment **do**
    Detect the start time t_{start} using ConvLSTM:
      $P_{\text{start}}(t) = \text{ConvLSTM}_{\text{start}}(X_t)$
    Retain frames from t_{start} to the current time step T:
      $X_{\text{truncated}} = x_{\text{start}}, x_{\text{start}+1}, \ldots, x_T$
    Predict activity using ConvLSTM on truncated sequence:
      $P_{\text{ongoing}}(t) = \text{ConvLSTM}_{\text{ongoing}}(X_{\text{truncated}})$
  end for
end procedure

Algorithm 1. Bi-Directional ConvLSTM-Based Action Recognition

Testing procedure

The action starting point detection network forecasts the beginning point tstart(t) for the ongoing sequence that terminates at the current time step during the testing phase. The prediction network then assesses the truncated sequence Xtruncated, generating a probability distribution Pongoing(t) for the continuous operation.

To summarize, our design consists of two specialized ConvLSTM networks: one for predicting ongoing activity and another for detecting action beginning points. During training, these networks function alone, and during testing, they are seamlessly merged. Real-time prediction of ongoing activities is made possible by the action beginning point detection network, which provides the prediction network with information on the initiation point. Effective early identification of ongoing actions in unsegmented data streams is made possible by this two-step procedure.

Results and discussions

To evaluate the performance of the proposed Bi-ConvLSTM architecture, we conducted several experiments using a dual NVIDIA Tesla P100 GPU setup with 3584 CUDA cores and a peak computational throughput of 18.7 TeraFLOPS. All experiments were implemented using the Darknet deep learning framework, in combination with Python 3.9 and TensorFlow 2.11.

The model was trained for 50 epochs using a batch size of 64. The Adam optimizer was employed with a learning rate of 0.0013, and categorical cross-entropy was used as the loss function for both the ConvLSTMstart and ConvLSTMongoing networks. To reduce overfitting, dropout regularization with a rate of 0.5 was applied to the recurrent units.

Input video sequences were resized to 128×128 pixels, and frames were sampled at a fixed rate of 30 frames per second (fps). During training, each input sequence was segmented into overlapping windows of 20 frames with a stride of 10 frames.

Both ConvLSTM networks used in the architecture consisted of a single ConvLSTM layer with 64 filters and a 3×3 kernel size, followed by ReLU activations and MaxPooling layers. The final classification layer employed a softmax activation function to produce a probability distribution over the activity classes.

This implementation setup ensures that the model can be trained efficiently while preserving both spatial and temporal information necessary for accurate early activity recognition.

In addition, we used the Temple dataset to evaluate the performance of several advanced deep learning models. These include Convolutional Neural Networks (CNN)[30], InceptionV3[31], VGG19[32], and hybrid architectures such as InceptionV3-LSTM[2533]. We also compared the performance of ConvLSTM[2], Bi-ConvLSTM[334], and our proposed improved Bi-ConvLSTM model. This comprehensive evaluation aims to determine the effectiveness of each model in classifying human activities in the dataset. To meet the specific objectives of our study, some model parameters were fine-tuned accordingly.

The evaluation considers a real-world situation in which the beginning of a continuous action is not known in advance, which makes the categorization task much more difficult. Several strategies can be used in these circumstances to guarantee precise categorization. Examining many possible action beginning points is a useful strategy that enhances prediction resilience and enables the model to reflect different temporal circumstances. By examining the activity from many temporal viewpoints, this method eventually produces a more accurate and dependable categorization result, guaranteeing that the model reaches the most certain conclusion.

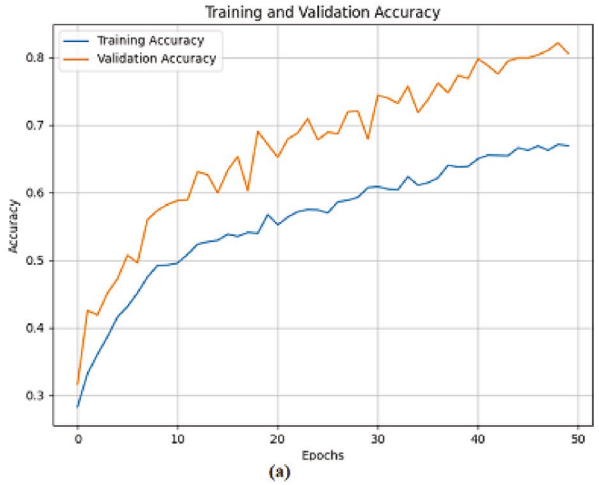
The accuracy performance of several models for classifying human activity is broken down in Table 3, which shows that the suggested Bi-ConvLSTM model performs noticeably better than the others. For example, the mean accuracies of CNN, InceptionV3, and VGG19 are 82.29%, 86.49%, and 79.67%, respectively. The Bi-ConvLSTM, on the other hand, exhibits exceptional accuracy and consistency with a mean accuracy of 89.52% and a low standard deviation of 0.08 as shown in Table 2. Thus, the performance of the Bi-ConvLSTM model is measured as follows: the standard deviation (σ) quantifies the departure of these accuracy values from the mean, and the mean accuracy (x) is the average of all accuracy values acquired over several runs (Fig. 4).

The formula for the mean accuracy and standard deviation is given in the equation

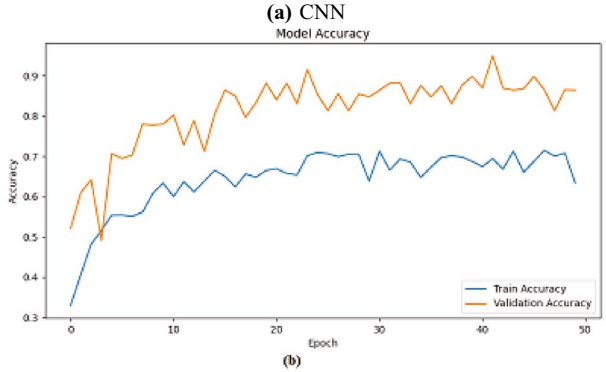
$$x = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

Model	Layers & Units	Accuracy (%)
CNN[36]	2 Conv Layers (16, 32 filters), Max-Pooling, Dropout, FC Layer (128 units)	82.29
InceptionV3[37]	InceptionV3 Layers, GlobalAveragePooling2D, Dense Layers	86.53
VGG19[38]	Frozen Layers + Custom Dense Layers with Dropout and Batch Normalization	79.74
InceptionV3-LSTM[39]	InceptionV3 (Frozen), GlobalAveragePooling2D, LSTM (128, 64 units)	83.54
ConvLSTM[2]	2 ConvLSTM2D Layers (32, 64 filters), Batch Normalization, Max-Pooling, FC Layers	84.53
Bi-ConvLSTM[340]	2 Bidirectional ConvLSTM2D Layers, Batch Normalization, Max-Pooling, Dense Layers	85.78
Proposed Bi-ConvLSTM Model	1 st Layer: Bidirectional ConvLSTM2D (32 filters), 2nd Layer: Bidirectional ConvLSTM2D (64 filters), Batch Normalization, Dense Layers	**89.54**

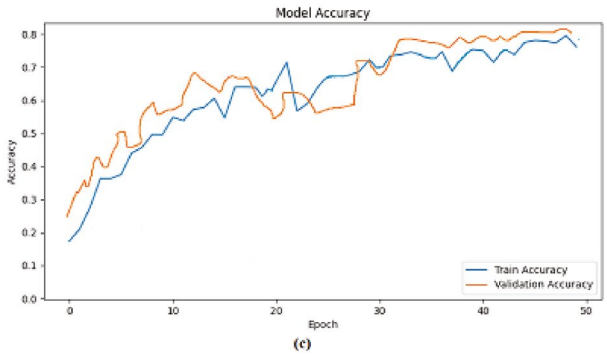
Table 2. Comparison of Human Activity Classification Models.



(a) CNN



(b) InceptionV3

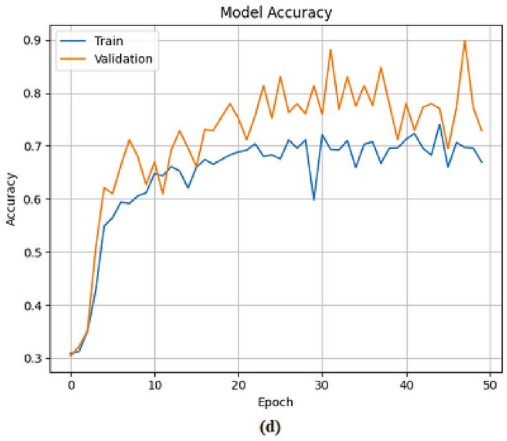


(c) VGG19

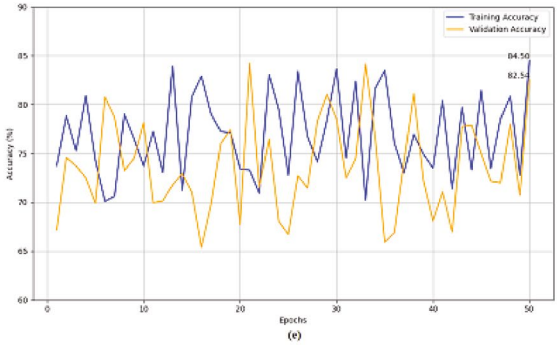
Fig. 4. Comparison of accuracy for CNN, InceptionV3, and VGG19 models. Comparison of test and validation accuracy across advanced models including InceptionV3_LSTM, ConvLSTM, Bi-ConvLSTM, and the proposed model.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

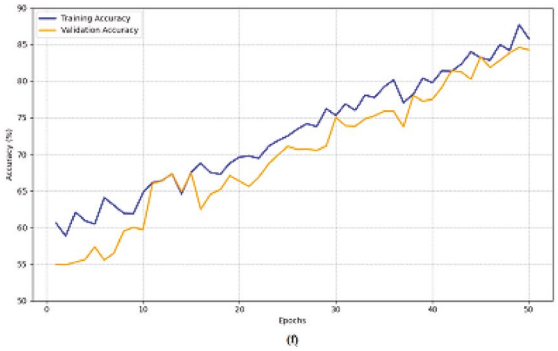
where n is the number of runs and xi stands for each accuracy value. The Bi-ConvLSTM model's considerably greater mean accuracy and smaller standard deviation demonstrate how well it performs and maintains stability in tasks involving the classification of human activities. These results are graphically depicted in Fig. 5, which highlights the Bi-ConvLSTM's notable performance advantage. Bi-ConvLSTM's dual-layer design offers a strong foundation for comprehending and forecasting complicated sequences, with distinct layers for determining beginning points and forecasting continuing actions. This method provides thorough insights from the input sequences by utilizing the advantages of convolutional integration and bidirectional processing. The model



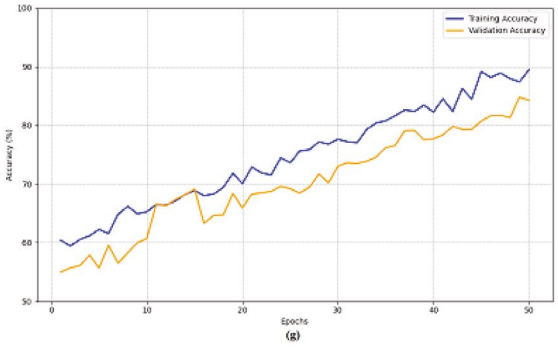
(d) InceptionV3_LSTM



(e) ConvLSTM



(f) Bi-ConvLSTM



(g) Proposed Bi-ConvLSTM

Fig. 4. (continued)

is especially well-suited for challenges requiring complicated sequences and action prediction because it can represent complex temporal dynamics and bidirectional relationships.

The main computational challenges of implementing the Bi-ConvLSTM framework include high memory usage and increased computational complexity due to the bidirectional nature and convolutional operations. The bidirectional processing requires maintaining hidden states for both forward and backward passes, which significantly increases memory requirements and computational load. Convolutional operations further add to the complexity by involving multiple filter applications and weight updates.

These challenges were addressed through several strategies: reducing input size by resizing frames to a fixed size of 224×224 pixels and using selective data augmentation techniques such as rotation, flipping, and brightness adjustment to lower memory requirements and computational overhead; optimizing the model architecture by tuning the number of hidden units and convolutional layers to balance complexity and performance; and employing mini-batch training along with GPU acceleration to handle large data volumes efficiently and reduce training time. These strategies significantly improved computational efficiency while maintaining high model accuracy, making the Bi-ConvLSTM framework suitable for real-world human activity recognition tasks in complex environments like temples.

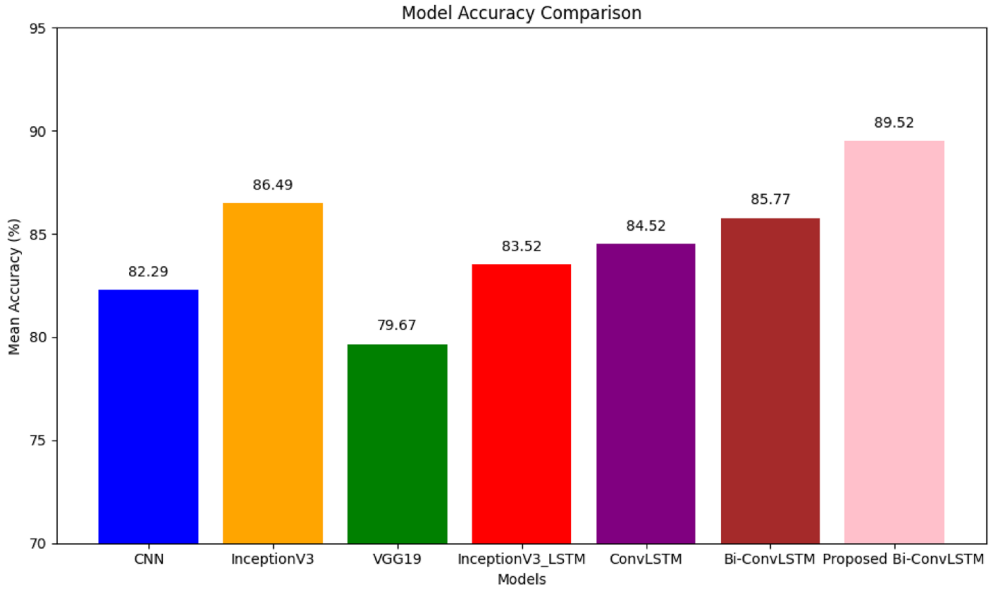


Fig. 5. Model comparisons with the proposed model.

Model	Accuracy Values(%)	Mean Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Standard Deviation
CNN	[82.00, 82.50, 82.40, 82.30, 82.29]	82.29	82.00	81.20	82.10	0.22
InceptionV3	[86.20, 86.60, 86.40, 86.70, 86.53]	86.49	86.20	86.30	86.25	0.22
VGG19	[79.50, 79.80, 79.70, 79.60, 79.74]	79.67	79.50	79.60	79.55	0.12
InceptionV3-LSTM	[83.40, 83.60, 83.50, 83.55, 83.54]	83.52	83.40	83.50	83.45	0.08
ConvLSTM	[84.40, 84.60, 84.55, 84.50, 84.53]	84.52	84.40	84.50	84.45	0.08
Bi-ConvLSTM	[85.70, 85.80, 85.75, 85.85, 85.78]	85.77	85.70	85.75	85.72	0.06
Proposed Bi-ConvLSTM	[89.40, 89.60, 89.50, 89.55, 89.54]	**89.52**	89.40	89.50	89.45	0.08

Table 3. Performance Comparison of Various Models for Human Activity Classification.

The Bi-ConvLSTM model consistently outperforms other models in terms of mean accuracy and variance stability because of its bidirectional processing capacity and the integration of convolutional layers with LSTMs. The model's bidirectional nature enables it to capture both past and future temporal relationships, facilitating comprehension of complicated sequences. Furthermore, the convolutional layers allow for successful feature extraction from input data, and the LSTM units manage long-term dependencies, improving the model's predictive ability. This combination design enables the model to generalize effectively across several data samples, resulting in high accuracy and low variation.

Conclusion and future scopes

In conclusion, the comparison study shown in Fig. 4 and Table 3 highlights the outstanding performance of the suggested Bi-ConvLSTM model for classifying human activities. Compared to CNN, InceptionV3, and VGG19, the Bi-ConvLSTM model performs better, with a noteworthy mean accuracy of 89.52% and a low standard deviation of 0.08. This significant accuracy gain and excellent consistency across several runs demonstrate how well the Bi-ConvLSTM performs in producing better outcomes. Its substantial potential to advance human activity categorization tasks is demonstrated by the model's capacity to provide both improved performance and stability. According to the results, the suggested Bi-ConvLSTM is a strong and dependable solution that significantly outperforms current approaches.

A number of approaches might be investigated in future studies to improve early recognition systems even further. Accuracy and resilience in a variety of settings may be enhanced by integrating multi-modal data, such as RGB and depth information. Furthermore, the model's usefulness may be improved by extending it to accommodate more intricate and diverse activity categories, such as interactions between several individuals. Additional advancements in early activity identification could come from further research into hybrid models that incorporate ConvLSTM with other cutting-edge methods like transformers or attention processes.

Data availability

Data is available from the corresponding author on reasonable request

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

Received: 1 February 2025; Accepted: 1 October 2025
Published online: 06 November 2025

References

1. Shenoy, A., & Thillaiarasu, N. "A survey on different computer vision based human activity recognition for surveillance applications." In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 1372-1376. IEEE, (2022).
2. Ashwin Shenoy, M. & Thillaiarasu, N. 'Enhancing Temple Surveillance through Human Activity Recognition: A Novel Dataset and YOLOv4-ConvLSTM Approach'. 11217 – 11232 (2022).
3. Wang, Boyu & Hoai, Minh. Back to the beginning: Starting point detection for early recognition of ongoing human actions. *Computer Vision and Image Understanding* **175**, 24–31 (2018).
4. Thillaiarasu, N. & Shenoy, Ashwin. "Enhancing Security Through Real-Time Classification of Normal and Abnormal Human Activities: A YOLOv7-SVM Hybrid Approach." IAENG International Journal of Computer Science 51, no. 8 (2024).
5. Hoai, M. & De la Torre, F. Max-margin early event detectors. *International Journal of Computer Vision* **107**, 191–202 (2014).
6. Ryoo, M. *Human activity prediction: Early recognition of ongoing activities from streaming videos* (In Proc, ICCV, 2011).
7. Ryoo, M. S., Fuchs, T. J., Xia, L., Aggarwal, J. K. & Matthies, L. Robotcentric activity prediction from first-person videos: What will they do to me? In International Conference on Human-Robot Interaction, (2015).
8. Ashwin Shenoy, M., Thillaiarasu, N. & Shenoy, Ashwin. "Enhancing Security Through Real-Time Classification of Normal and Abnormal Human Activities: A YOLOv7-SVM Hybrid Approach," IAENG International Journal of Computer Science. **51**(8), 1027-1034, (2024).
9. Vondrick, C., Pirsiavash, H., & Torralba, A. "Anticipating the future by watching unlabeled video." arXiv preprint arxiv.org/abs/1504.08023 **2**(2), (2015).
10. Hu, J.F., Zheng, W.S., Ma, L., Wang, G. & Lai, J. Real-time rgb-d activity prediction by soft regres- sion, in: Proceedings of the European Conference on Computer Vision. (2016).
11. Li, Y., Lan, C., Xing, J., Zeng,W., Yuan, C. & Liu, J. Online human action detection using joint classification-regression recurrent neural networks, in: Proceedings of the European Conference on Com- puter Vision. (2016).
12. Ryoo, M. S., Fuchs, T. J., Xia, L., Aggarwal, J. K. & Matthies, L. Robotcentric activity prediction from first-person videos: What will they do to me? In (2015).
13. Cao, Y., et al. Recognize human activities from partially observed videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013).
14. Hoai, M., & Zisserman, A. "Improving human action recognition using score distribution and ranking." In Asian conference on computer vision, 3-20. Cham: Springer International Publishing, (2014).
15. Hoai, M., & Zisserman, A. "Talking heads: Detecting humans and recognizing their interactions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 875-882. (2014).
16. Simonyan, K., & Zisserman, A. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014).
17. Sadegh Aliakbarian, M., et al. "Encouraging lstms to anticipate actions very early." In Proceedings of the IEEE International Conference on Computer Vision, 280-289. (2017).
18. Koppula, H. S. & Saxena, A. *Anticipating human activities for reactive robotic response* (In Proc, IROS, 2013).
19. Wang, Z., et al. "Probabilistic modeling of human movements for intention inference." Proceedings of robotics: Science and systems, VIII (2012).
20. Wang, Jack M., Fleet, David J. & Hertzmann, Aaron. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence* **30**(2), 283–298 (2007).
21. Cao, S., & Nevatia, R. "Forecasting human pose and motion with multibody dynamic model." In 2015 IEEE Winter Conference on Applications of Computer Vision, 191-198. IEEE, (2015).
22. Pavlovic, V., Rehg, J. M., & MacCormick, J. "Learning switching linear models of human motion." Advances in neural information processing systems 13 (2000).
23. Wang, Y., & Hoai, M. Improving human action recognition by non-action classification. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2698–2707, (2016).
24. Pavlovic, V., Rehg, J. M. & MacCormick, J. Learning switching linear models of human motion. *Advances in neural information processing systems* **13** (2000).
25. Sadegh A., et al. Encouraging lstms to anticipate actions very early. In 2017 IEEE International Conference on Computer Vision (ICCV), 280–289, (2017).
26. Wang, Y., Zhou, W., Zhang, Q., Zhu, X., & Li, H. Weighted multi-region convolutional neural network for action recognition with low-latency online prediction. In 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW), 1–6, (2018).
27. Tran, V., Balasubramanian, N., & Hoai, M. Progressive knowledge distillation for early action recognition. In 2021 IEEE International Conference on Image Processing (ICIP), 2583–2587, (2021).
28. Tran, V., Wang, Y., Zhang, Z., & Hoai, M. Knowledge distillation for human action anticipa- tion. In 2021 IEEE International Conference on Image Processing (ICIP), 2518–2522, (2021).
29. Wang, Boyu & Hoai, Minh. Back to the beginning: Starting point detection for early recognition of ongoing human actions. *Computer Vision and Image Understanding* **175**, 24–31 (2018).
30. Tuhin Kumar B. "A Brief Introduction to Human Activity Recognition Using Deep Learning", Futuristic Trends in Computing Technologies and Data Sciences Volume 3 Book 3, IIP Series, Volume 3, 1-16, e-ISBN: 978-93-6252-901-5, (2024). https://doi.org/10.58532/V3BFCT3P1CH1
31. SaiRamesh, L., Dhanalakshmi, B., & Selvakumar, K. Human Activity Recognition Through Images Using a Deep Learning Approach. (2024). https://doi.org/10.21203/rs.3.rs-4443695/v1
32. Gandhi, V. Human Activities Recognition Using Machine Learning and Artificial Initialization. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/cseit2410276 (2024).
33. Rajanidi, S., et al. "Towards Real-Time Human Activity Recognition: A Machine Learning Perspective," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 1-9, https://doi.org/10.1109/ICCCNT61001.2024.10724335. (2024).
34. Anbazhagan, K., Swamy, G., Janani, R. & Farakte, A. "Deep Learning based Human Activity Recognition in Smart Home," 2024 4th International Conference on Data Engineering and Communication Systems (ICDECS), Bangalore, India, 1-6, https://doi.org/10.1109/ICDECS59733.2023.10502527. (2024)
35. Lu, Z. "Deep Learning-Based Human Activity Recognition Algorithms: A Comparative Study," 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 1041-1046, https://doi.org/10.1109/ICIPCA59209.2023.10257913. (2023).

36. Raj, Ravi & Kos, Andrzej. An improved human activity recognition technique based on convolutional neural network. *Sci. Reports* **13**(1), 22581 (2023).
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016).
38. Simonyan, K., & Zisserman, A. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
39. Tran, D., Bourdev, L., Maji, S., & Papanikolopoulos, N. "Learning Spatiotemporal Features with 3D Convolutional Networks." IEEE International Conference on Computer Vision (ICCV). (2015).
40. Arif, S., & Wang, J. "Bidirectional LSTM with saliency-aware 3D-CNN features for human action recognition." Journal of Engineering Research. **9**(3), (2021).

Author contributions
A. S. M.: Conceptualization, Methodology, Experimentation, Data Collection, Writing – Original Draft, Supervision. T. N: Methodology, Data Analysis, Writing - Review & Editing, Validation. S.S.: Software Implementation, Data Preprocessing, Visualization, Formal Analysis. S.K.S: Literature Review, Data Preprocessing, Visualization, Writing - Review & Editing, Manuscript Formatting. All authors have read and approved the final version of the manuscript.

Funding
No funding was received from any financial organization to conduct this research.

Declarations

Ethical approval
All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the Institutional Ethics Committee (IEC) of NMAM Institute of Technology (NMAMIT), Nitte, Karnataka, India. Informed consent was obtained from all subjects and/or their legal guardian(s).

Conflicts of interest
The authors declare that they have no known financial or non-financial competing interests in any material discussed in this paper.

Additional information
Correspondence and requests for materials should be addressed to M.A.S.