



## HARGAN: Generative Adversarial Network Based Deep Learning Framework for Efficient Recognition of Human Actions from Surveillance Videos

Boddupally JANAIHAH<sup>1\*</sup>, Suresh PABBOJU<sup>2</sup>

<sup>1</sup>Research Scholar Osmania university, Department of CSE Hyderabad India

\* **Corresponding Author Email:** janaiahmvsr@gmail.com - **ORCID:** 0009-0009-3456-4086

<sup>2</sup>Professor Chaitanya Bharathi institute of Technology Department of IT, Hyderabad India.

**Email:** [plpsuresh@gmail.com](mailto:plpsuresh@gmail.com) - **ORCID:** 0000-0001-8309-7562

### Article Info:

DOI: 10.22399/ijcesen.587

Received : 04 November 2024

Accepted : 10 November 2024

### Keywords :

Human Action Recognition  
Artificial Intelligence,  
Deep Learning,  
Generative Adversarial Network,  
CNN,  
LSTM.

### Abstract:

Analyzing public surveillance videos has become an important research area as it is linked to different real-world applications. Video Analytics for human action recognition is given significance due to its utility. However, it is very challenging to analyze live-streaming videos to identify human actions across the frames in the video. The literature showed that Convolutional Neural Networks (CNNs) are among computer vision applications' most popular deep learning algorithms. Another important observation is that Generative Adversarial Network (GAN) architecture with deep learning has the potential to leverage effectiveness in applications using computer vision. Inspired by this finding, we created a GAN-based framework (called HARGAN) in this research for human activity identification from surveillance films. The framework exploits a retrained deep learning model known as ResNet50 and convolutional LSTM for better performance in action recognition. Our framework has two critical functionalities: feature learning and human action recognition. The ResNet50 model achieves the former, while the GAN-based convolutional LSTM model achieves the latter. We proposed an algorithm called the Generative Adversarial Approach for Human Action Recognition (GAA-HAR) to realize the framework. We used a benchmark dataset known as UCF50, which is extensively used in studies on human action identification. Based on our experimental findings, the suggested framework performs better than the current baseline models like CNN, LSTM, and convolutional LSTM, with the highest accuracy of 97.73%. Our framework can be used in video analytics applications linked to large-scale public surveillance.

## 1. Introduction

Acquiring massive labeled video datasets is one of the significant obstacles in video activity recognition. Because videos can vary in viewpoint, size, and look, annotating them can be costly and time-consuming. This implies that semi-supervised methods for action recognition in videos are required. Using deep networks to develop a feature representation of films without activity labels but with frame temporal order as a form of "weak supervision" is one such method [1,2]. To construct good questions with correct ordering for the deep network's input, as well as to decide on sampling strategies and associated video encoding techniques such as dynamic pictures [3], supervision is still necessary for this approach. Generative models, like the recently released Generative Adversarial

Networks (GANs) [4], use an adversarial process to approximate large dimensional probability distributions, such as those of actual photographs, without needing costly labeling. Typically, GANs are used to start with random noise and learn an image data distribution. Two networks are involved in adversarial learning in GANs: a generator and discriminator networks. Two different types of inputs are used to train the discriminator network: random noise and samples taken from high-dimensional data sources like photographs. Its objective is to differentiate between produced and actual samples. The discriminator's output is sent into the generator network to produce "better" samples. This minimax game aims to get the discriminator to converge to a state where it cannot tell the difference between produced and actual data.

We suggest learning a feature representation of the discriminator by using it to solely distinguish between a natural and produced sample. We train a discriminator network using the GAN setup and utilize the learned discriminator representation as the "initialized weight." Next, refine that discriminator using labelled video datasets like UCF101 [5]. While several studies have conducted small-scale tests recently [6], to our knowledge, no comprehensive research has specifically examined all the architecture/hyperparameter configurations that may produce high performance on multiple datasets (including HMDB51, where we perform well) using simply appearance information from the video. This unsupervised pre-training phase, which uses just appearance information, produces action recognition performance that is competitive with state-of-the-art without requiring human feature engineering, video frame encoding, or looking for the optimum video frame sampling approach. From the literature, it was observed that methods for deep learning, like Convolutional Neural Networks (CNN), are widely used for computer vision applications. Another important observation is that Generative Adversarial Network (GAN) architecture with deep learning has the potential to leverage effectiveness in applications using computer vision. This feature inspired us to present in this research a GAN-based system for recognizing human actions from surveillance footage, called HARGAN. The following are our contributions to this publication.

1. We presented the HARGAN framework, a GAN-based system for recognizing human actions from surveillance videos. The framework exploits a pre-trained deep learning model known as ResNet50 and convolutional LSTM for better performance in action recognition.
2. To realize the framework, we proposed an algorithm called the Generative Adversarial Approach for Human Action Recognition (GAA-HAR).
3. We developed a prototype to evaluate our framework and algorithm for efficient human action recognition, and the state of the art is contrasted.

The rest of the paper is organized this way: Section 2 offers a current literature assessment of earlier research on video analytics for human action detection. Section 3 presents the suggested architecture, workings, and underlying algorithm for recognizing human actions. Section 4 shows experiment findings using the UCF50 dataset and compares them to the most advanced models. Our research is concluded in Section 5, which offers potential directions for further study in this field.

## 2. Related work

There are numerous deep learning approaches for recognizing human actions in videos. Kambala and Jonnadula [1], based on deep neural networks, the OPA-PPAR technique preserves accurate activity identification while improving privacy. Empirical research demonstrates its advantages. Wu et al. [2] maximized trade-offs between privacy and usefulness, a unique approach that tackles privacy-preserving action recognition in deep learning. Validations through experiments demonstrate efficacy. Liu et al. [3] presented CIASA, the first adversarial approach using GCNs to guarantee spatial integrity and temporal coherence in skeleton-based action recognition. Sun et al. [4] preserved high usefulness to identification tasks, a novel framework that produces human action pictures with verifiable privacy guarantees. Ahuja et al. [5] explained using Doppler radar in the millimeter wave range to identify activities. Human-computer interaction is advanced by converting movies to synthetic radar data using a software pipeline that addresses privacy issues and data shortages. Hou et al. [6] suggested a super-resolution GAN that improves accuracy over current techniques for activity identification in extremely low-resolution videos. Jiang et al. [7] employed body-part segmentation rather than skeletons; the BPA-GAN for human motion transfer increases training efficiency and reconstruction quality. Gedamu et al. [8] expanded the training set's view range, suggesting in arbitrary-view human action recognition challenges, the Two-Branch Novel-View Generation technique improves recognition performance. Pikramenos et al. [9] explained the bias problem in motion recognition datasets and suggested enhancing classifier generalization across camera angles by combining pose extraction with domain adaption. The outcomes demonstrate efficacy, especially in terms of cross-view alterations. Jimale et al. [10] suggested a better CGAN design to increase sample quality and model stability for sensor-based activity identification. Unsupervised and semi-supervised settings may be investigated in future research. Tan et al. [11] surpassed current methods and proposed a Bi-LSTM-based model for older people living alone that preserves their privacy while recognizing activities. Yan et al. [12] presented a method that obscures target data using image segmentation masks to enable privacy-preserving Human Action Recognition (HAR). Imran et al. [13] improved accuracy and protected privacy by employing deep learning to bridge the gap in identifying aggressive behavior in real-time monitoring. Liang et al. [14] addressed sub-action sharing issues using a

segmental architecture and CPPCR for multimodal human action recognition. Rajput et al. [15] suggested using position-based superpixel modification to obfuscate color and depth data to enhance security and reduce data overhead in a secure human action recognition method. Wang et al. [16] explored characteristics that are used instead of image restoration within a coded aperture camera system without a lens that is intended for privacy-preserving human activity detection. Chaudhary et al. [17], with the growing usage of camera sensors, issues in computer vision include data size and user privacy. We propose a Pose Guided Dynamic Image network for human activity recognition to overcome these challenges effectively. Hao et al. [18] drew to assisted living using Human Activity Recognition (HAR). The robustness of recognition is enhanced via a novel Wi-Fi-based neural network (WiNN). Wu et al. [19] included crowded backdrops and perspective variance in video-based human action detection, which is vital in many disciplines. Deep learning methods developed recently are promising. Obaidi et al. [20] presented a technique that uses temporal salience modeling and HOG-S features to anonymize films while maintaining action recognition usefulness visually. Bach et al. [21] the necessity of real-time perioperative condition assessment is highlighted by technological advancements in hospitals—privacy concerns prompt identity de-identification solutions. A successful prototype based on YOLO v3 recognizes and de-identifies problematic regions in OR pictures, promoting privacy-aware processing and further research developments. Dai et al. [22] investigated action recognition using simulation in low smart-room settings for temporal and spatial resolution. The results show that recognition is achievable and highlight the need for spatial resolution. Perera et al. [23] presented MOD20, a multi-class action recognition dataset consisting of 2324 films from YouTube and a drone. Liu et al. [24] suggested a technique for video action identification that protects privacy by utilizing C3D networks, compressed sensing, and the AdSRC algorithm. Duta et al. [25] presented ST-VLAD, a spatiotemporal encoding technique that improves action recognition performance on complicated datasets. Wang et al. [26] suggested a safe aggregation and edge computing-based collaboration architecture for privacy-preserving HAR model training. The results show high precision and a suitable training duration. Data poisoning detection and consensus protocol upgrades are examples of future advancements. Angelini et al. [27] presented a 2D pose-based HAR technique called ActionXPose that uses OpenPose in settings similar to CCTV. Verified using the ISLD dataset, it exhibits cutting-edge accuracy and

resilience. Integration with internet surveillance, RGB integration, and generalization are future development goals. Roshtkhari and Levine [28] presented a robust hierarchical codebook model for action recognition and video matching that is resistant to deformations and variations in scale. Long-term behavior knowledge and enhanced spatial and temporal connection modeling are the main goals of future research. Farrajota et al. [29] suggested merging low-level body joint data with high-level ConvNet properties to create a system for action recognition, securing cutting-edge outcomes on the FLIC, LSP, and UCF Sports datasets, forthcoming research endeavors to broaden its scope and integrate motion data. Yang et al. [30] investigated ways for monitoring human behavior while protecting privacy, classifying them into signal, algorithm, and system levels. For scholars and practitioners, it offers perspectives and recommendations for the future. Xu et al. [31] highlighted effective optical flow feature extraction and spatio-temporal integration while introducing a quick network for human activity identification. Reaching optimal accuracy, it rapidly outperforms earlier models on a range of datasets. Ramanathan et al. [32] described obstacles to action recognition, including view shifts and occlusion, and evaluated the performance of current techniques. It draws attention to areas needing further investigation and datasets for future work. Silva and Marana [33], with competitive accuracy rates, the method extracts spatiotemporal characteristics for action identification by converting 2D poses into  $(\theta, \rho)$  parameter space. Dhamsania and Ratanpara [34], for Content-Based Video Retrieval to be important in computer vision research, human activity recognition is necessary. Many approaches encounter difficulties. Abdelbaky and Aly [35] suggested a human action recognition system that works better than existing methods using PCANet with motion energy templates. Agahian et al. [36] created a brand-new framework that uses 3D skeleton data, posture representation, PCA, Fisher Vector encoding, and ELM to recognize human actions. Although little has been studied, Gutoski et al. [37] state that open-set human action recognition (HAR) is essential. We present a deep metric learning model, TI3D, which achieves better results on UCF101. Kar et al. [38] presented AdaScan, an adaptive temporal pooling technique that enhances baseline approaches based on accepted standards for identifying human actions. Yang et al. [39] combined CNN and RNN; SCNN can directly extract spatial-temporal information from video frames, greatly enhancing action detection ability. Pareek and Thakkar et al. [40] examined ML and DL approaches for HAR, encompassing action

recognition algorithms, datasets, and applications between 2011 and 2019. In the literature, it has been noted that deep learning methods such as Convolutional Neural Networks (CNN) are extensively utilized in computer vision applications. It is also important to note that the architecture of Generative Adversarial Networks (GAN) in conjunction with deep learning has the potential to significantly enhance effectiveness in computer vision applications.

### 3. Preliminaries

GAN is extensively utilized for many real-time applications, such as picture synthesis and human action identification, to name a few. The discriminator (D) and generator (G) networks

comprise this system. Both are utilized based on deep neural networks with implicit functionality. Figure 1 illustrates the GAN's structure. G captures the dynamics of the distribution from actual data samples.  $G(z)$  represents the collected data, while  $p_g(z)$  represents its distribution. GAN seeks to match the distribution of the training sample, represented as  $p_r(x)$ , to  $p_g(z)$ .  $G(z)$  and actual data  $x$  are the inputs supplied to D. D's result is the likelihood that its input comes from a proper distribution. G creates a pseudo-sample, indicated as  $G(z)$ , with Gaussian noise by mapping random variables, denoted as  $z$ , to neural networks. While G is being trained, D has fixed parameters. Before data created by G is sent to D, it is marked as fraudulent. The result of the  $D(G(z))$  is used to represent D and the discrepancy between it and the sample label is calculated.

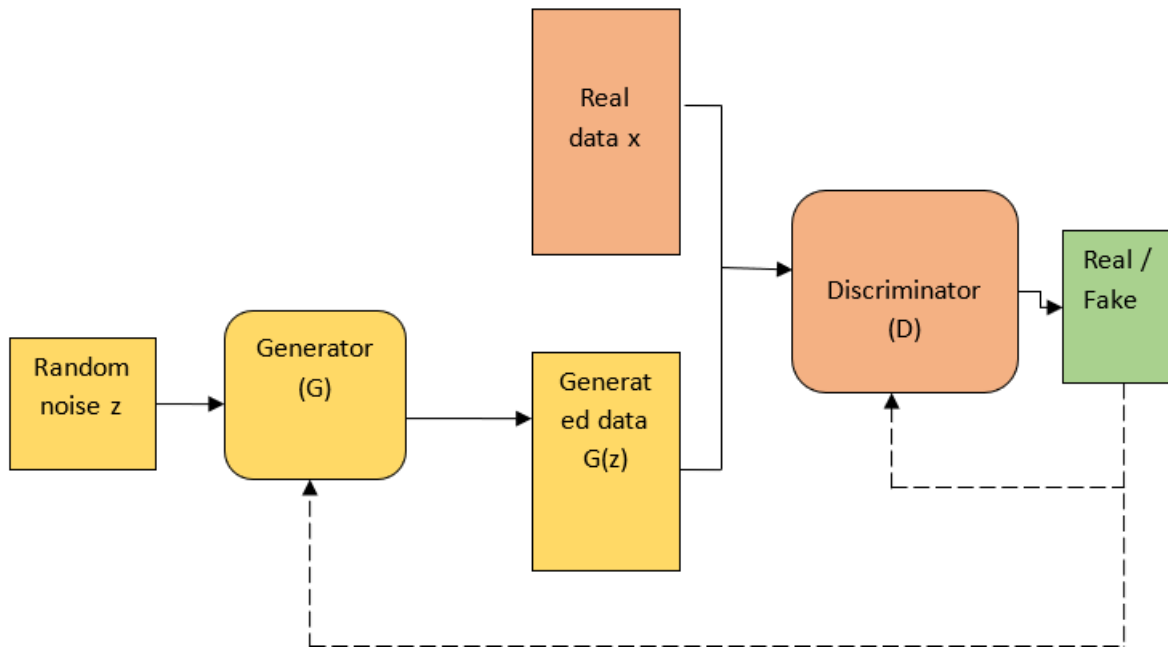


Figure 1: A typical GAN framework

The parameters of G are updated using the backpropagation approach. The discriminator wants to know if the input comes from a reliable sample and gives the proper feedback so that G's settings may be adjusted. The output of D approaches 1 and 0 in response to whether the input is an accurate sample  $x$ . D receives  $x$  as well as  $G(z)$  as input. A standard GAN framework uses the mini-max game between two players to compute the loss function of two neural networks that compete with one another in a zero-sum game situation. D's loss function may be found in Equation 1.

$$V(D, \theta^{(D)}) = -E_{x \sim p_r(x)}[\log D(x)] - E_{z \sim p_g(z)}[\log(1 - D(g(z)))] \tag{1}$$

Similarly, the G has its loss function as in Eq. 2.

$$V(G, \theta^{(G)}) = E_{z \sim p_g(z)}[\log(1 - D(g(z)))] \tag{2}$$

Each participant in the game has a loss function. During the process, D aims to maximize  $V^{(D)}(\theta^{(D)}, \theta^{(G)})$  and G updates  $\theta^{(D)}$  and  $\theta^{(G)}$  to maximize  $V^{(G)}(\theta^{(D)}, \theta^{(G)})$ . The players' loss functions are dependent on parameters. Nash equilibrium must be reached for a player to stop training and update the settings of another player.

$$\text{Min}_{\theta^{(D)}} \max_{\theta^{(G)}} V(D, G) = E_{x \sim p_r(x)}[\log D(x)] + E_{z \sim p_g(z)}[\log(1 - D(g(z)))] \tag{3}$$

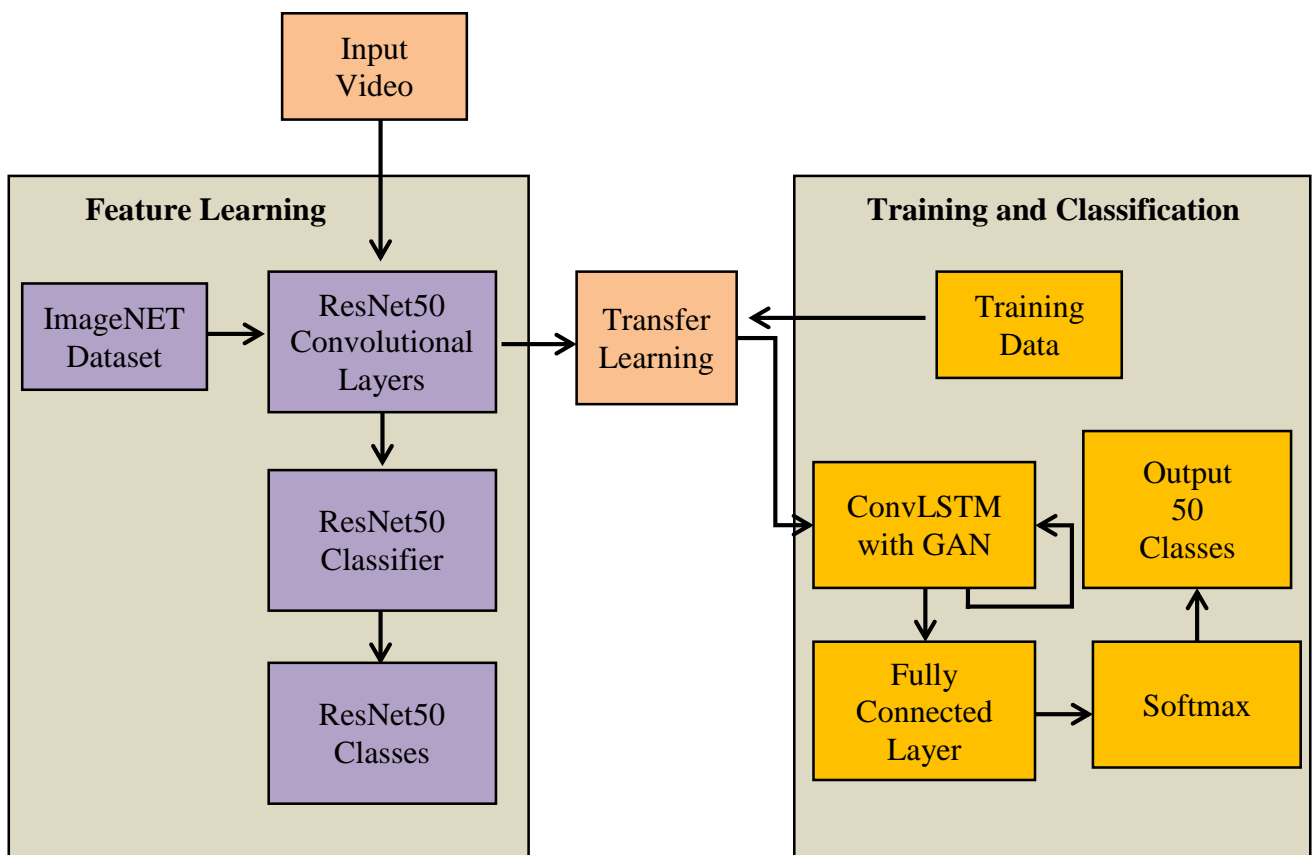
Equation 3 is the loss function of a GAN, which is modeled as a mini-max optimization problem. As a result, D creates an objective function that is as big as feasible using actual data as input.  $D(G(z))$  is the

output of D, and its objective is to ensure that the output is near 1. After enough practice, the game reaches a Nash equilibrium.

#### 4. Proposed system

We proposed a GAN-based framework for efficiently recognizing human actions from surveillance videos. The proposed GAN-based architecture has provisions for feature engineering and also action recognition. Figure 2 displays a summary of the suggested system. A deep learning model known as ResNet50 is used for feature learning, which is essential for action recognition later. The other part of the proposed framework

exploits the GAN model to leverage the action recognition procedure. The action recognition part of the proposal system is based on the convolutional LSTM model. ConvNet is a deep neural network with a focus on image identification. It uses a network of neurons to detect objects by simulating the structure of the human visual cortex. Since video data is sequential and each frame impacts the one after it, this research employs LSTM to tackle the problem. A recurrent network called LSTM gives the subsequent node feedback from the preceding node. The suggested model uses an adversarial loss function instead of the traditional cross-entropy, avoiding clashes between classes and regenerating output and feedback to the input.



**Figure 2:** Proposed GAN-based system for human action recognition from videos

The ResNet50 model is vital in extracting features from a given input video. This deep learning model is pre-trained with the ImageNet dataset. However, the model is retrained with the training dataset comprising labeled surveillance videos. Features obtained from the feature learning module of ResNet50 are used in the training and classification module. The proposed system exploits GAN architecture, which is made up of a convolutional LSTM model. The GAN architecture performs its iterative process with a generator and discriminator involved in action recognition in surveillance videos.

The outcome of the GAN model is given to a completely linked layer, a softmax layer, and finally, multi-class classification.

##### 4.1 Feature Learning

Although there are several approaches to extracting characteristics from video data, the most successful approach is transfer learning. The suggested approach uses the ResNet50 model, a pre-trained residual network that was pre-trained using the ImageNet dataset. The last wholly linked layer is removed and integrated with the UCF50 training set



to accomplish feature learning using the ResNet50 architecture [12]. ResNet50, moreover, is not the output layer and classifier needed in this case since the data is not steady. A deep residual learning model appropriate for medical image processing is called ResNet50. There are fifty layers deep. This particular type of CNN is highly in-depth because of its creative use of skip connections. Convolution and identity blocks are the two shortcut modules used in the ResNet50 implementation. The latter lacks a convolution layer at the shortcut, but the former

contains one. Block-out dimensions in a convolution are greater than input dimensions. of contrast, the output dimensions of the identity block match the input dimensions. For both blocks, there are 1x1 convolution layers at the beginning and the end. This type of method, called bottleneck design, lowers the parameters without sacrificing performance. Specific deep shortcut modules are eliminated, and further classification layers are introduced in the empirical investigation.

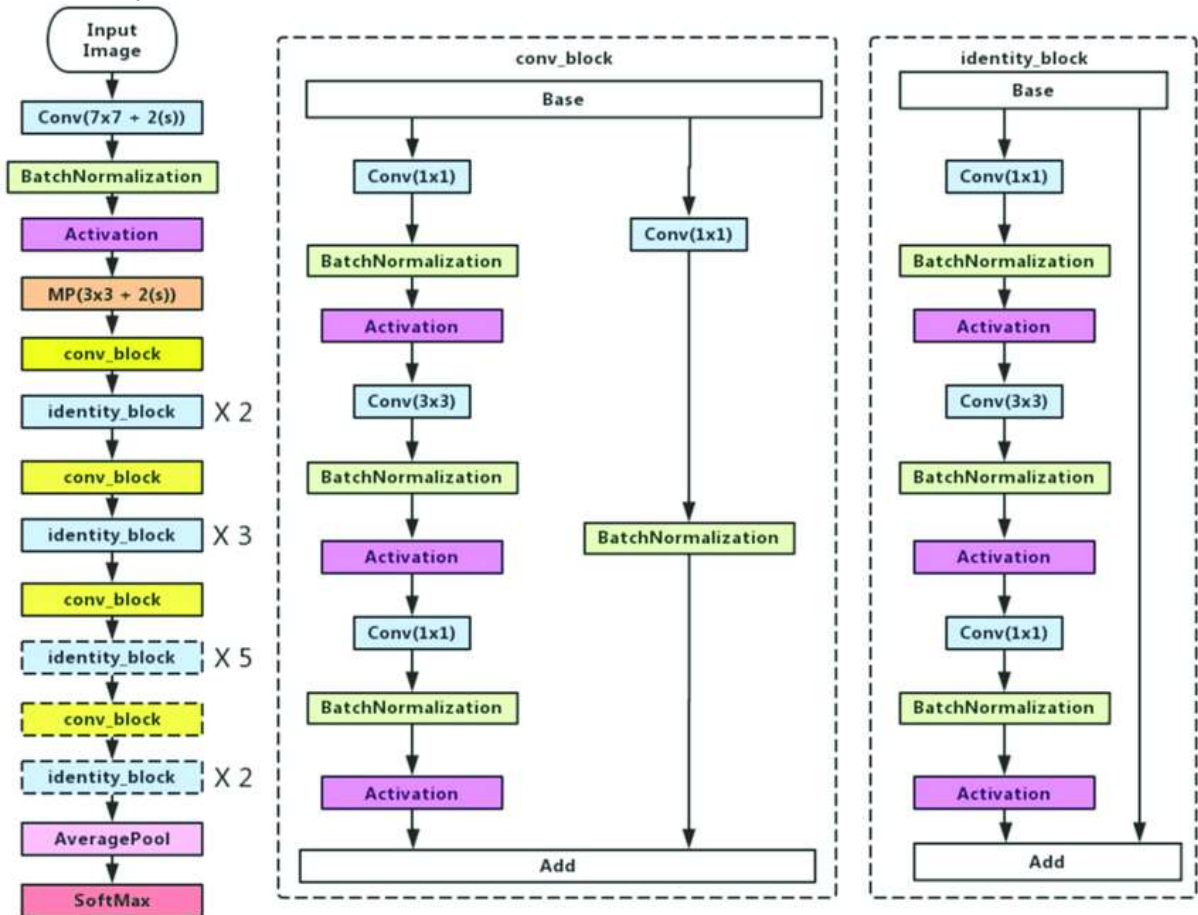


Figure 3: Overview of ResNet50 architecture

Figure 3 displays the redesigned ResNet50 architecture with the identity and convolution blocks in the proper and middle schematics. Due to requirements, the ResNet50 architecture's dotted-line blocks represent the ones eliminated for our study. While the identity block does not alter the input's dimensions, the convolution block does. Equation 1 represents a building component used in the residual learning process.

$$y = F(x, \{W_i\}) + x \tag{1}$$

where  $x$  and  $y$  represent the input and output vectors and express the residual mapping process as  $F(x, \{W_i\})$ . Eq. 1 also expresses shortcut connections without additional parameters and avoids computational complexity. The optimized strategy improves performance. If the dimensions of  $x$  and  $F$

in Eq. 1 are different, a linear projection is carried out as stated in Eq. 2.

$$y = F(x, \{W_i\}) + W_s x \tag{2}$$

The linear projection, denoted by  $W_s$ , should only be utilized when the input and output dimensions match.  $F(x, \{W_i\})$  can represent more than one convolutional layer, and the residual function is adjustable. Convolutional layers are composed of a number of filters (matrix-based) that use the activation function to create feature maps after converting the picture through convolution operations. To produce a desired range of output from a given input, one uses the fundamental mathematical function known as an activation function. Rectified linear units (ReLU), hyperbolic

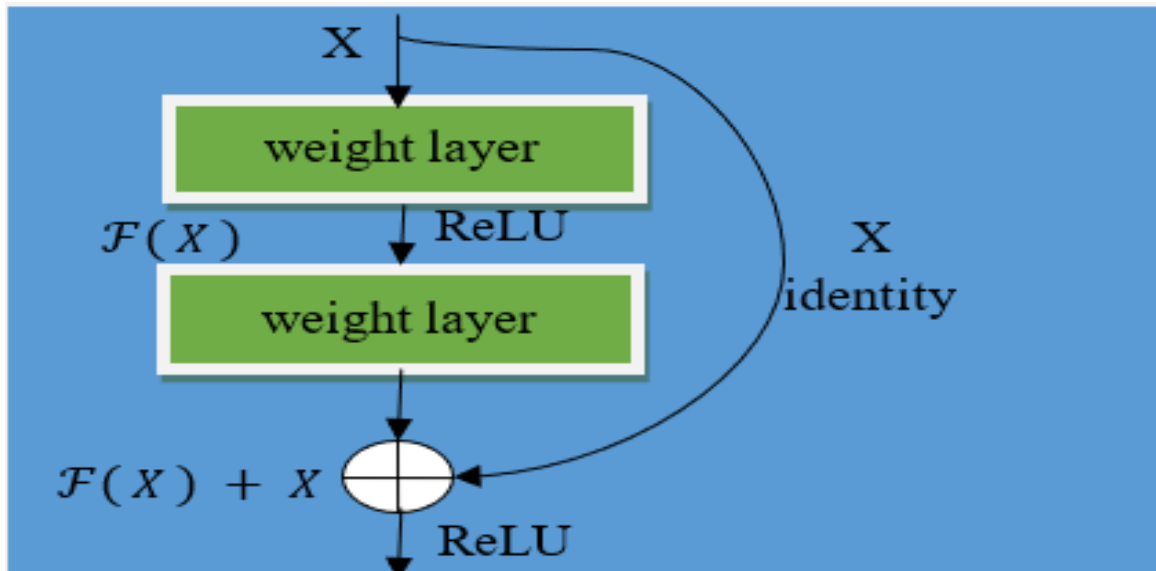


Figure 4: Illustrates residual network in ResNet model

tangent Tanh, and logistic sigmoid are a few examples of the various activation functions that are available. Numerous researchers have demonstrated that ReLU performs better in deep neural networks than other activation functions [1,11]. The experiment utilizes ReLU as an activation function. Note that the number of convolution filters is always equal to the number of feature maps generated.

Getting around one or more levels is the main idea behind the ResNet. As seen in Figure 4, Residual functionals specify the layers concerning input but exclude connections because identity functions are complex to learn. Equation 3 expresses the function in question.

$$F(x) = x \quad (3)$$

Theoretically, improving the identity mapping might make achieving residual to zero easier. What happens if the subsequent layer's spatial dimension differs from the layer's output before it after the skip connection? It is expressed as in Eq. 4.

$$y = f(x, W_i) + W_s x \quad (4)$$

Then, various strategies have been put forth to upgrade the picture dimensions, such as projecting the dimensions using a 1x1 convolution or padding the zeros. Because high-performance computing power is costly and time-consuming, transfer learning was developed to provide researchers access to pre-trained models, which increased training efficacy. More processing power is needed to train a deeper network. The most profound network at the moment is the ResNET50 model. ConvNet differs from other neural networks in that

it performs all its operations in a two-dimensional plan utilizing the convolutional and pooling layers.

#### 4.2 Action Recognition

The intricate action recognition process involves learning from available data and predicting upcoming observations. In this case, the neighboring pixels are combined using the pooling layer to minimize the size. Mean pooling or max pooling might be used. While mean or average pooling produces a smooth extraction, max pooling produces an extraction with sharp edges. In this experiment, the borders of human body parts are detected using max pooling. The last pooling layer of the ResNet learning section's features is employed to execute classification operations in this paper's classification part. The classification layer comprises many stages, including the output layer, softmax, fully connected layer, and ConvLSTM-based GAN. The GAN architecture involved in the proposed system is presented in Figure 5.

Let  $V$  be a collection of videos, where  $n$  denotes the dataset's number of films and  $V = \{v_1, \dots, v_n\}$ . A configurable number of frames make up each video. We train the GAN model using all the frames in the training set of movies from two challenging video activity datasets without any label information. ConvLSTM-based GAN exploits the flatten technique used in this step to turn the inputs from the feature learning section's pooling layer into a single row. Alternatively, it converts the input into a single dimension. Using the batch normalization technique, the input is normalized before learning non-linearity. To learn the linear characteristics of the data, a linear network receives the flatten layer's output. Benchmark average and standard deviation are computed in Eq. 5 and Eq. 6.

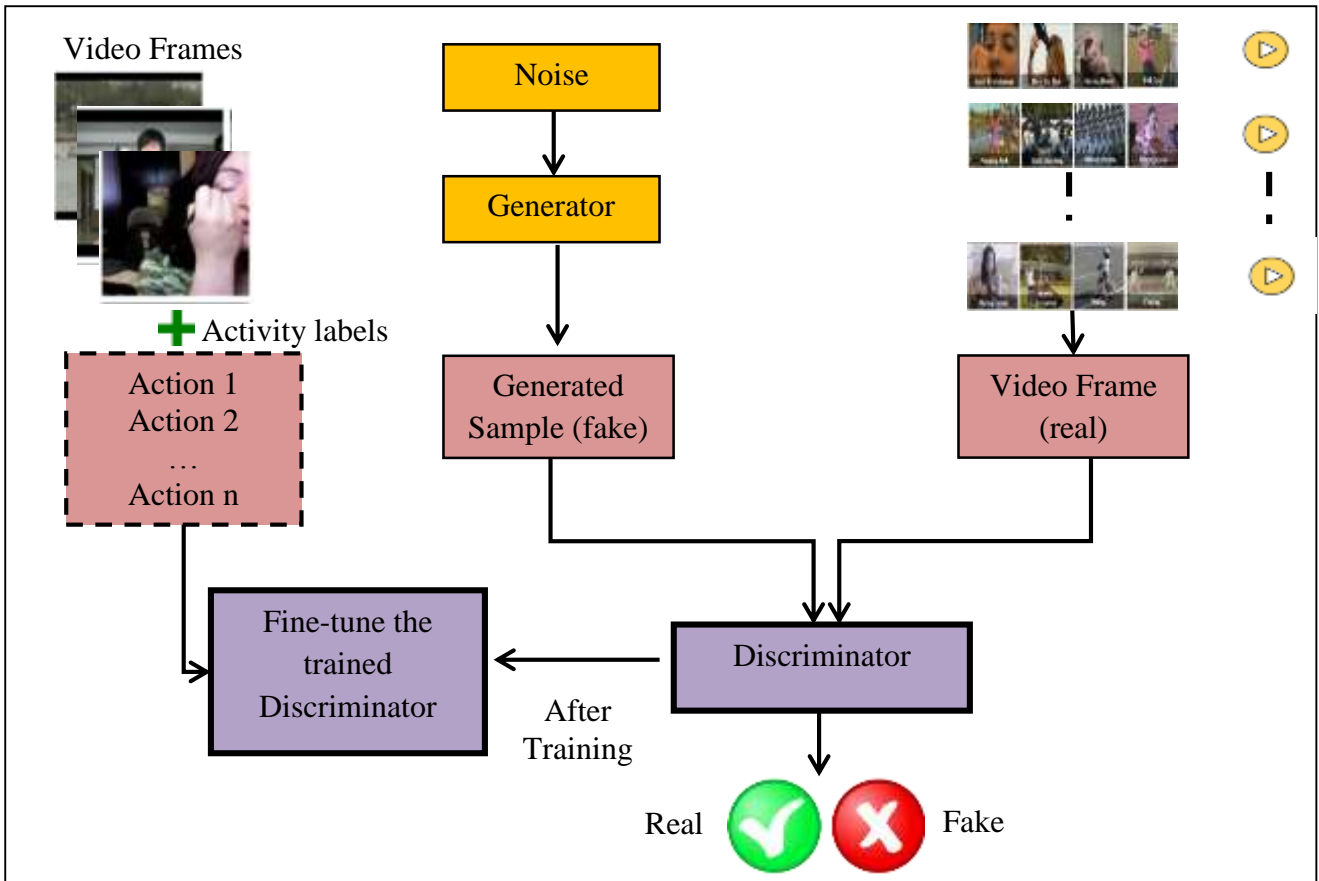


Figure 5: GAN architecture used in the proposed system

$$\mu = \frac{1}{k} \sum x_i \quad (5)$$

$$\sigma^2 = \frac{1}{k} \sum_{i=1}^k (x - \mu)^2 \quad (6)$$

Then, the normalization process results in the expression in Eq. 7.

$$x_i = \frac{x_i - \mu}{\sigma_2 + \epsilon}, \forall i = 1, \dots, k \quad (7)$$

Remember that these parameters may be learned, allowing for optimization using techniques like gradient descent, as expressed in Eq. 8.

$$x_i = Yx_i + \beta \quad (8)$$

It has been demonstrated that batch normalization accelerates training by a factor of 11.4. Then, the non-linear function is used to discover the data's non-linearity components. ReLU is used in the proposed model as a non-learning function, often known as an activation function. We use an activation feature, which involves a fully connected network because the output is non-linear. Since the output of the convolutional layer is also non-linear, I am using a fully connected layer in the classification layer. In the fully connected layer, each output dimension depends on every input dimension. Softmax is another kind of activation function, often called the logistic function. It is typically used for categorical distribution and ranges from 0 to 1. The last layer of the model receives the output probabilities and uses them to determine the

specific category of the input video. The anticipated outcome can be seen in the last classification step. The output layer has a step count of "c," where "c" is the number of categories that must be classified.

### 4.3 Dataset and Loss Function

LSTM and CNN are used in the experiment as part of GAN, with extra assistance from transfer learning associated with the ResNet50 model. The dataset utilized in this investigation is UCF50 [41]. The dataset has 50 action categories with diversified videos. This data set is a benchmark dataset as it has a number of variations in terms of illumination conditions, background, viewpoint, and object scale besides object appearance and camera motion. Given the size of the UFC101 Dataset, manual feature creation is nearly impossible. ConvLSTM is used to execute training against the freshly supplied data. Once the pre trained ResNET50 model is housed in the training block, further data is introduced. ConvLSTM is a CNN and LSTM combination used for training operations and feature extraction from sequential data. Choosing the loss function is the most important but often overlooked choice. It is the function that is being reduced using a variety of techniques. Many techniques exist; one popular one is cross-entropy. Nevertheless, the suggested approach uses generative adversarial networks,



commonly known as adversarial loss functions (GANs). The generator and discriminator are the two adversarial networks that make up the adversarial loss mechanism, as expressed in Eq. 9 and Eq. 10.

$$L_{\epsilon} = E \left[ \log \left( D_{\phi} \left( I, \epsilon_{\theta(I)} \right) \right) \right] \quad (9)$$

$$L_D = E \left[ \log \left( D_{\phi} \left( I, \Gamma_M \right) \right) + \log \left( 1 - D_{\phi} \left( I, \epsilon_{\theta(I)} \right) \right) \right] \quad (10)$$

Here, D serves as the discriminant. We want to find disparities between the estimator's output and manually segmented cells. To solve the min-max issue, the expression in Eq. 11 is used.

$$\min_{\phi} \max_{\theta} \left\{ E \left[ \log \left( D_{\phi} \left( I, \Gamma_M \right) \right) + \log \left( 1 - D_{\phi} \left( I, \epsilon_{\theta(I)} \right) \right) \right] \right\} \quad (11)$$

ConvLSTM extracts several characteristics by utilizing both linear and non-linear functions. Before the non-linear function, batch normalization is added to increase performance. The optimum activation function for picture training is ReLU, which is a non-linear function in this method. Average pooling is applied after the non-linear function to lower the dimension, and the fully linked network receives the outcome of this procedure. Lastly, using the softmax, the approach describes each class probabilistically for the final step.

#### 4.4 Proposed Algorithm

Our suggested algorithm, the Generative Adversarial Approach for Human Action Recognition (GAA-HAR), realizes the framework. We used a benchmark dataset known as UCF50, which is widely used in human action recognition research (Algorithm 1).

**Algorithm:** Generative Adversarial Approach for Human Action Recognition (GAA-HAR)  
**Input:** UCF50 dataset D (surveillance videos)  
**Output:** Results of human action recognition R, performance statistics P

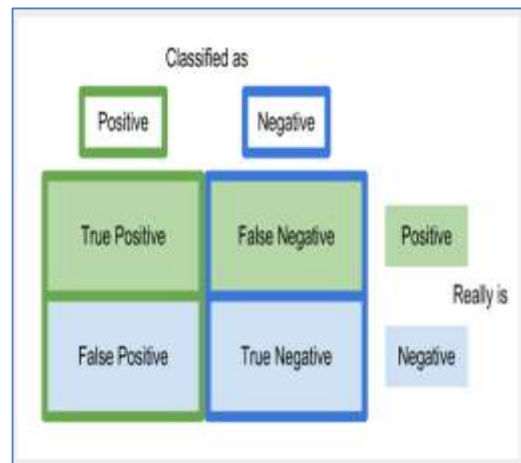
1. Begin
2. (T1, T2) ← SplitData(D)
3. Train ResNet50 with ImageNet
4. Retrain ResNet50 with T1
5. featureMaps ← FeatureLearningUsingResNet50(T1)
6. For each featureMap in featureMaps
7. Generator function
8. Discriminator function
9. End For
10. optimizedFeatureMaps ← GANWithConvLSTM(featureMaps)
11. R ← RecognizeActions(FC, T2)
12. P ← Evaluation(R, ground truth)
13. Display R
14. Display P
15. End

**Algorithm 1:** Generative Adversarial Approach for Human Action Recognition

As seen in Algorithm 1, it recognizes desired actions from provided films using the UCF50 dataset as input. The provided dataset is split into training (T1) and test (T2) sets. Using the ImageNet dataset, a deep learning model called ResNet50 is developed. For further performance improvement, the ResNet50 model is retrained with the training data obtained from the UCF50 dataset with the help of convey knowledge. There's a cyclical procedure in which the generative adversarial network takes feature maps generated by ResNet50 as input and generates optimized feature maps. In the process, the generator and discriminator are functional until the feature maps are converged for improved performance. Then, the final feature maps are used for multi-class classification; the fully connected layer comes first, then the softmax layer, resulting in action recognition from all test videos.

#### 4.5 Evaluation Method

Figure 6 theoretically illustrates a confusion matrix. It illustrates four scenarios where the suggested system may identify a particular test sample. True Positive (TP) cases occur when an actual event occurs in the sample, and the suggested algorithm recognizes it as the same event. True Negative (TN) cases are those in which the suggested method correctly identifies the absence of any specified event in the supplied sample. False Positive (FP) occurs when an event occurs in the provided sample, and the suggested algorithm interprets it as a distinct event. In the event that the suggested algorithm recognizes a different event, even if there isn't a specific event in the provided sample, we refer to this situation as False Negative (FN). Various performance measures are created and applied to assess the suggested system based on the confusion matrix and the abovementioned situations. Another popular statistic for assessing performance is accuracy, which is given in Eq. 10.



**Figure 6:** Confusion matrix

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Accuracy is measured When comparing the number of accurate predictions made by an anomaly detection model to all of its predictions. The resultant value of this statistic ranges from 0.0 to 1.0, signifying the lowest and maximum performance, respectively.

## 5. Experimental results

This section shows our experiment's findings regarding human action recognition using the benchmark data set UCF50 [42]. Observations compare ground-truth actions under the proposed framework's predictions. Our experimental findings were also contrasted with the latest baseline models available regarding accuracy, including CNN, LSTM, and convolutional LSTM. CNN, which stands for Convolutional Neural Network, is a type of deep neural network primarily used for analyzing visual imagery. CNNs are specifically designed to automatically and adaptively learn spatial hierarchies of features directly from image data. They are composed of multiple layers of small neuron collections that process portions of the input image, known as receptive fields. On the other hand, LSTM, short for Long Short-Term Memory, is a type of recurrent neural network (RNN) architecture developed to address the vanishing gradient problem encountered by traditional RNNs. LSTMs excel at learning long-term dependencies in data sequences

by integrating a memory cell, input gate, forget gate, and output gate. The memory cell enables LSTMs to retain information over extended periods, while the gates regulate the flow of information in and out of the cell. Figure 7 gives an example from the UCF50 dataset, which contains several videos widely used for human action recognition research. As shown in Figure 8, video frames from the UCF50 data set are presented in different categories. The sample actions presented include rowing, juggling, and applying eye makeup. Action recognition results of the proposed framework are provided in Figure 9. It shows different sample frames, their ground truth labels, and the proposed framework's predicted labels. Figure 10 displays the experimental findings about the number of epochs and the accuracy of our GAN-based model. The model's accuracy increases with epoch count, and its validation accuracy gradually increases until convergence. At epoch 56, the model converged with the highest performance. Figure 11 displays the experiment's outcomes in terms of the loss of our GAN-based model against the number of epochs. As the number of epochs increases, model loss and its validation loss gradually decrease until convergence. At epoch 56, the model converged with the least possible loss. Findings from an experiment regarding the accuracy of different human action recognition models are presented in Table 1. Human action recognition performance of different deep learning models are contrasted, and the outcomes are displayed as Figure 12 illustrates.

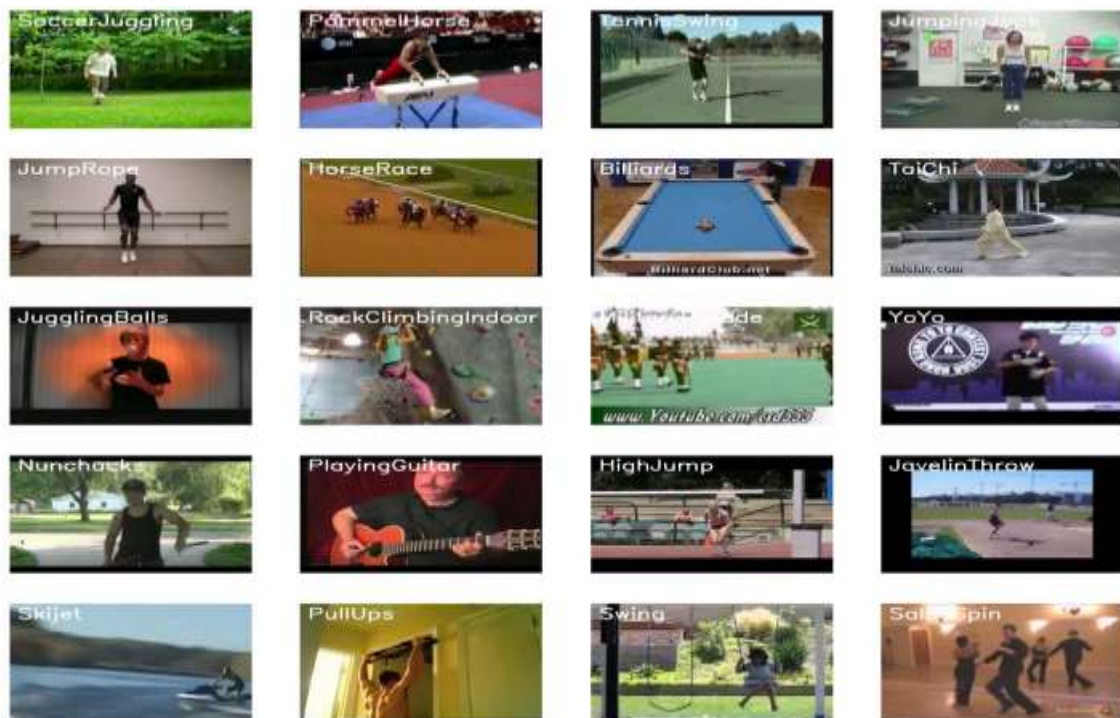


Figure 7: An excerpt from the UCF50 dataset



Figure 8: Sample frames from the UCF50 dataset







Input Frame	Ground Truth Label	Predicted Label
	Robbery	Robbery
	Fighting	Fighting
	Playing Guitar	Playing Guitar
	High Jump	High Jump
	Pushups	Pushups
	Walking with Dog	Walking with Dog

Figure 9: Action recognition results of the proposed system



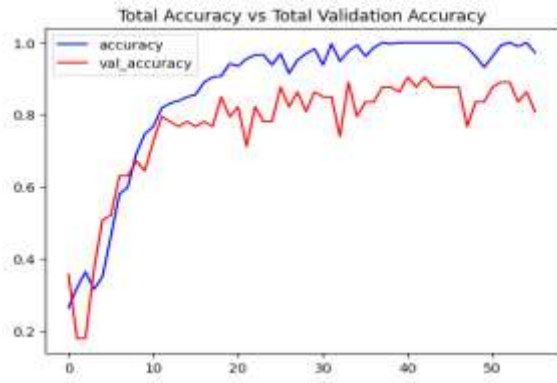


Figure 10: Accuracy of the proposed model

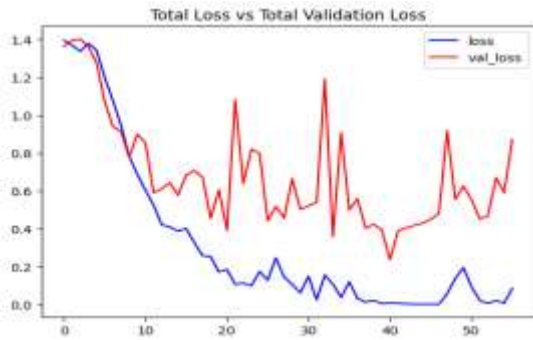


Figure 11: Loss of proposed model

Table 1: Performance comparison among different models

HAR Model	Accuracy (%)
CNN	87.34
LSTM	90.54
ConvLSTM	95.23
GAA-HAR (Proposed)	97.73

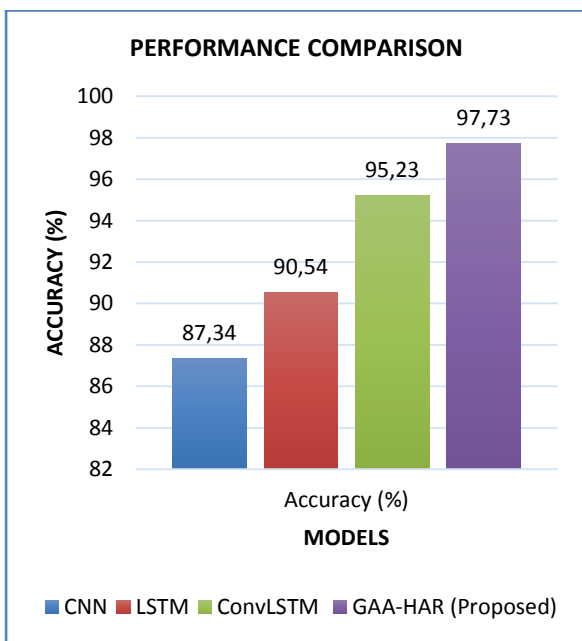


Figure 12: Performance comparison among models

The proposed GAN-based approach is compared with existing deep learning models like CNN, LSTM, and convolutional LSTM. Each model showed a different level of accuracy due to the difference in their modulus operandi. The baseline CNN model could achieve 87.34% accuracy, while the LSTM model achieved 90.54% accuracy, which is better than the baseline CNN model. The convolutional LSTM model is a hybrid approach based on CNN and LSTM models and could improve human action recognition performance significantly. The model achieved 95.23% accuracy, which is better than the accuracy exhibited by CNN and LSTM models. However, the proposed GAN-based model outperformed all existing models with the highest accuracy of 97.73%. Therefore, the suggested GAN-based method is found to be a better candidate for video analytics in public surveillance applications.

### 6. Discussion

For various reasons, human action recognition in surveillance videos plays an essential role in the contemporary era. Video surveillance is commonly used in public places and other vital locations to provide safety measures. Surveillance videos consist of frames with image data; computer vision and artificial intelligence techniques are widely used to process this data. However, deep learning methods are insufficient in video data analytics. This paper proposes a deep learning-based framework, which utilizes GAN architecture and a convolutional LSTM model for efficient detection of human actions in surveillance videos. The framework involves feature learning, training, and classification. The feature learning process utilizes a pre-trained deep learning model called ResNet50. This model is retrained with the UCF-50 dataset using transfer learning to improve the feature learning process. The GAN-based architecture used in this research helps leverage diversity in features, leading to efficient detection of human actions. The classification process of the framework results in multi-class classification. GAN architecture and a specially designed feature learning process make the proposed framework more efficient. However, the proposed framework has certain limitations, as discussed in section 5.1.

#### 6.1 Limitations of the Study

The proposed framework has certain limitations. The framework only uses one dataset, which may not be sufficient to generalize the findings in this research. Additionally, the framework does not consider the discrimination of abnormal activities from everyday activities. Another significant

limitation is that the framework recognizes actions but does not provide privacy protection for the identity of human beings.

## 7. Conclusion and future work

In this research, we presented a GAN-based architecture for surveillance video-based human activity identification. The framework exploits a pre-trained deep learning model known as ResNet50 and convolutional LSTM for better performance in action recognition. Our framework has two critical functionalities: feature learning and human action recognition. The ResNet50 model achieves the former, while the latter is achieved by the GAN-based convolutional LSTM model. The method we suggested for human action recognition is the Generative Adversarial Approach (GAA-HAR) to realize the framework. We used a benchmark dataset known as UCF50, which is extensively used in studies on human action recognition. Our testing results showed that the suggested framework performs better than baseline models currently in use, such as CNN, LSTM, and convolutional LSTM, with the highest accuracy of 97.73%. Our framework can be used in video analytics applications linked to large-scale public surveillance. Our mythology has certain limitations. We evaluated the proposed framework with only one data set, UCF50. However, generalizing findings with one specific data set may cause threats to the validity of the proposed method. Another limitation is that the proposed system does not include privacy-preserving phenomena throughout the identification of human action. In the future, we intend to improve our framework by overcoming these limitations. Artificial Intelligence is an important method and thus there have been many different works done on it [43-55].

### Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by **Bodupally Janaiah, Suresh**

**Pabboju** The first draft of the manuscript was written by **S Bodupally Janaiah** all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

- [1] V.K. Kambala, H. Jonnadula, (2022). Privacy preserving human activity recognition framework using an optimized prediction algorithm, *IAES International Journal of Artificial Intelligence (IJ-AD)*,11(1);254-264.
- [2] Z. Wu, H. Wang, Z. Wang, H. Jin, Z. Wang,(2022). Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset, *IEEE Trans Pattern Anal Mach Intell* 44(4):2126-2139. doi: 10.1109/TPAMI.2020.3026709
- [3] J. Liu, N. Akhtar, A. Mian, (2020). Adversarial Attack on Skeleton-Based Human Action Recognition, *IEEE Transactions on Neural Networks and Learning Systems*, 1–14. doi:10.1109/tnnls.2020.3043002
- [4] M. Sun, Q. Wang, Z. Liu, (2020). Human Action Image Generation with Differential Privacy, *IEEE International Conference on Multimedia and Expo (ICME)*, 1-6. doi:10.1109/icme46284.2020.910276
- [5] K. Ahuja, Y. Jiang, M. Goel, C. Harrison,(2021). Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Vid2Doppler:* 1-10. <https://doi.org/10.1145/3411764.3>
- [6] M. Hou, S. Liu, J. Zhou, Y. Zhang, Z. Feng, (2021). Extreme Low-Resolution Activity Recognition Using a Super-Resolution-Oriented Generative Adversarial Network, *Micromachines*, 12(6);1-15. <https://doi.org/10.3390/mi1206067>
- [7] J. Jiang, G. Li, S. Wu, H. Zhang, Y. Nie, (2021). BPA-GAN: Human motion transfer using body-part-aware generative adversarial networks, *Graphical Models*, 1151-10. doi:10.1016/j.gmod.2021.101107
- [8] K. Gedamu, Y. Ji, Y. Yang, L. Gao, H.T. Shen, (2021). Arbitrary-view human action recognition via novel-view action generation, *Pattern Recognition*, 118;1-9. doi:10.1016/j.patcog.2021.108043
- [9] G. Pikramenos, E. Mathe, E. Vali, I. Vernikos, A. Papadakis, E. Spyrou, P. Mylonas, (2020) An adversarial semi-supervised approach for action recognition from pose information, *Neural*



- Computing and Applications*, 1-15. doi:10.1007/s00521-020-05162-5
- [10] A.O. JIMALE, M.H.M. NOOR, (2022). Fully Connected Generative Adversarial Network for Human Activity Recognition, *IEEE Access*, 1-10.
- [11] T.H. Tan, L. Badarch, W.X. Zeng, M. Gochoo, (2021). Binary Sensors-Based Privacy-Preserved Activity Recognition of Elderly Living Alone Using an RNN, *MDPI* 1-18.
- [12] J. Yan, F. Angelini, S.M. Naqvi (2020), Image Segmentation Based Privacy-Preserving Human Action Recognition for Anomaly Detection, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1-5. doi:10.1109/icassp40776.2020.9054
- [13] J. Imran, B. Raman, A.S. Rajput, Robust, (2020). Efficient and privacy-preserving violent activity recognition in videos *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, doi:10.1145/3341105.3373942
- [14] C. Liang, D. Liu, L. Qi, L. Guan, (2020). Multi-Modal Human Action Recognition With Sub-Action Exploiting and Class-Privacy Preserved Collaborative Representation Learning, *IEEE Access*, 8 39920–39933. doi:10.1109/access.2020.2976496
- [15] A.S. Rajput, B. Raman, J. Imran, (2020), Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN, *Expert Systems with Applications* 113349 (2020) 1-15. doi:10.1016/j.eswa.2020.113349
- [16] Z.W. Wang, V. Vineet, F. Pittaluga, S.N. Sinha, O. Cossairt, S.B. Kang, (2019). Privacy-Preserving Action Recognition Using Coded Aperture Videos, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:10.1109/cvprw.2019.00007
- [17] S. Chaudhary, A. Dudhane, P. Patil, S. Murala, (2019), Pose Guided Dynamic Image Network for Human Action Recognition in Person Centric Videos, *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* 1-8. doi:10.1109/avss.2019.8909835
- [18] Y. Hao, Z. Shi, Y. Liu, (2020) A Wireless-Vision Dataset for Privacy Preserving Human Activity Recognition, *Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*, 1-9. doi:10.1109/mcna50957.2020.926428
- [19] D. Wu, N. Sharma, M. Blumenstein, (2017) Recent advances in video-based human action recognition using deep learning: A review, *International Joint Conference on Neural Networks (IJCNN)*, 1-8. doi:10.1109/ijcnn.2017.7966210
- [20] S.A. OBAIDI, H.A. KHAF AJI, C. ABHAYARATNE, (2020). Modeling Temporal Visual Saliency for Human Action Recognition Enabled Visual Anonymity Preservation, *IEEE Access*, 8;1-19.
- [21] T.N. Bach, D. Junger, C. Curio, O. Burgert, (2022) Towards Human Action Recognition during Surgeries using De-identified Video Data, *Current Directions in Biomedical Engineering*, 8(1);10 -112.
- [22] J. Dai, J. Wu, B. Saghafi, J. Konrad, P. Ishwar, (2015). Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1-9. doi:10.1109/cvprw.2015.7301356
- [23] A.G. Perera, Y.W. Law, T.T. Ogunwa, J. Chahl, (2020). A Multiviewpoint Outdoor Dataset for Human Action Recognition, *IEEE Transactions on Human-Machine Systems*, 1–9. doi:10.1109/thms.2020.2971958
- [24] J. Liu, R. Zhang, G. Han, N. Sun, S. Kwong, (2021) , Video action recognition with visual privacy protection based on compressed sensing, *Journal of Systems Architecture* 113;1-14. doi:10.1016/j.sysarc.2020.101882
- [25] I.C. Duta, B. Ionescu, K. Aizawa, N. Sebe, (2016) Spatio-Temporal VLAD Encoding for Human Action Recognition in Videos, *Lecture Notes in Computer Science*, 365–378. doi:10.1007/978-3-319-51811-4\_30
- [26] L. Wang, C. Zhao, K. Zhao, B. Zhang, S. Jing, Zhenxi, (2022). Privacy-Preserving Collaborative Computation for Human Activity Recognition, *Hindawi Security and Communication Networks*, 1-8.
- [27] F. Angelini, Z. Fu, Y. Long, L. Shao, S.M. Naqvi, (2019) 2D Pose-based Real-time Human Action Recognition with Occlusion-handling, *IEEE Transactions on Multimedia*, 1–14. doi:10.1109/tmm.2019.2944745
- [28] J.M. Roshtkhari, M.D. Levine, (2013). Human activity recognition in videos using a single example, *Image and Vision Computing*, 31(11);864–876. doi:10.1016/j.imavis.2013.08.005
- [29] M. Farrajota, J.M.F. Rodrigues, J.M.H. du Buf, (2018) Human action recognition in videos with articulated pose information by deep networks, *Pattern Analysis and Applications*, 1-12. doi:10.1007/s10044-018-0727-y
- [30] Y. Yanga, P. Hua, J. Shenb, H. Chengc, Z.A. Xiulo, (2024). Elsevier, Privacy-preserving human activity sensing: A survey, 1-35.
- [31] J. Xu, R. Song, H. Wei, J. Guo, Y. Zhou, X. Huang, (2021), A fast human action recognition network based on spatio-temporal features, *Neurocomputing* 441;350–358. doi:10.1016/j.neucom.2020.04.150
- [32] M. Ramanathan, W.Y. Yau, E.K. Teoh, (2014), Human Action Recognition With Video Data: Research and Evaluation Challenges, *IEEE Transactions on Human-Machine Systems* 44(5);650–663. doi:10.1109/thms.2014.2325871
- [33] M.V.D. Silva, A.N. Marana, (2020), Human action recognition in videos based on spatiotemporal features and bag-of-poses, *Applied Soft Computing*. 95;106513 doi:10.1016/j.asoc.2020.106513
- [34] C.J. Dhamsania, T.V. Ratanpara, (2016), A survey on Human action recognition from videos, *Online International Conference on Green Engineering and Technologies (IC-GET)* 1-5. doi:10.1109/get.2016.7916717

- [35] A. Abdelbaky, S. Aly, (2020), Human action recognition using short-time motion energy template images and PCANet features, *Neural Computing and Applications*. 32;12561–12574 doi:10.1007/s00521-020-04712-1
- [36] S. Agahian, F. Negin, C. Köse, (2019), An efficient human action recognition framework with pose-based spatiotemporal features, *Engineering Science and Technology, an International Journal*. doi:10.1016/j.jestch.2019.04.014
- [37] M. Gutoski, A.E. Lazzaretti, H.S. Lopes, (2020) Deep metric learning for open-set human action recognition in videos, *Neural Computing and Applications*, 33(4);1207–1220. <https://doi.org/10.1007/s00521-020-05009-z>
- [38] A. Kar, N. Rai, K. Sikka, G. Sharma, (2017). Adaptive Scan Pooling in Deep Convolutional Neural Networks for Human Action Recognition in Videos, *AdaScan*: 1-10. doi:10.1109/cvpr.2017.604
- [39] H. Yang, C. Yuan, J. Xing, W. Hu, (2017) SCNN: Sequential convolutional neural network for human action recognition in videos, *IEEE International Conference on Image Processing (ICIP)*, 1-5. doi:10.1109/icip.2017.8296302
- [40] P. Pareek, A. Thakkar, (2020), A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications, *Artificial Intelligence Review* 1-64. doi:10.1007/s10462-020-09904-8
- [41] D. Wu, N. Sharma, M. Blumenstein, (2017) , Recent advances in video-based human action recognition using deep learning: A review, *International Joint Conference on Neural Networks (IJCNN)* 1-8. doi:10.1109/ijcnn.2017.7966210
- [42] UCF50 - Action Recognition Data Set. Retrieved from <https://www.crcv.ucf.edu/data/UCF50.php>
- [43]M. Husain Bathushaw, & S. Nagasundaram. (2024). The Role of Blockchain and AI in Fortifying Cybersecurity for Healthcare Systems. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1120-1129. <https://doi.org/10.22399/ijcesen.596>
- [44] AY, S. (2024). Vehicle Detection And Vehicle Tracking Applications On Traffic Video Surveillance Systems: A systematic literature review. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1059-1068. <https://doi.org/10.22399/ijcesen.629>
- [45]Rama Lakshmi BOYAPATI, & Radhika YALAVARTHI. (2024). RESNET-53 for Extraction of Alzheimer’s Features Using Enhanced Learning Models. *International Journal of Computational and Experimental Science and Engineering*, 10(4);879-889. <https://doi.org/10.22399/ijcesen.519>
- [46]Sheela Margaret D, Elangovan N, Sriram M, & Vedha Balaji. (2024). The Effect of Customer Satisfaction on Use Continuance in Bank Chatbot Service. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1069-1077. <https://doi.org/10.22399/ijcesen.410>
- [47]jaber, khalid, Lafi, M., Alkhatib, A. A., AbedAlghafer, A. K., Abdul Jawad, M., & Ahmad, A. Q. (2024). Comparative Study for Virtual Personal Assistants (VPA) and State-of-the-Art Speech Recognition Technology. *International Journal of Computational and Experimental Science and Engineering*, 10(3);427-433. <https://doi.org/10.22399/ijcesen.383>
- [48]Güven, M. (2024). A Comprehensive Review of Large Language Models in Cyber Security. *International Journal of Computational and Experimental Science and Engineering*, 10(3);507-516. <https://doi.org/10.22399/ijcesen.469>
- [49]M, V., V, J., K, A., Kalakoti, G., & Nithila, E. (2024). Explainable AI for Transparent MRI Segmentation: Deep Learning and Visual Attribution in Clinical Decision Support. *International Journal of Computational and Experimental Science and Engineering*, 10(4);575-584. <https://doi.org/10.22399/ijcesen.479>
- [50]ÖZNAÇAR, T., & ERGENE, N. (2024). A Machine Learning Approach to Early Detection and Malignancy Prediction in Breast Cancer. *International Journal of Computational and Experimental Science and Engineering*, 10(4);911-917 <https://doi.org/10.22399/ijcesen.516>
- [51]Venkatraman Umbalacheri Ramasamy. (2024). Overview of Anomaly Detection Techniques across Different Domains: A Systematic Review. *International Journal of Computational and Experimental Science and Engineering*, 10(4);898-910. <https://doi.org/10.22399/ijcesen.522>
- [52]Türkmen, G., Sezen, A., & Şengül, G. (2024). Comparative Analysis of Programming Languages Utilized in Artificial Intelligence Applications: Features, Performance, and Suitability. *International Journal of Computational and Experimental Science and Engineering*, 10(3);461-469. <https://doi.org/10.22399/ijcesen.342>
- [53]Jafar Ismail, R., Samar Jaafar Ismael, Dr. Sara Raouf Muhamad Amin, Wassan Adnan Hashim, & Israa Tahseen Ali. (2024). Survey of Multiple Destination Route Discovery Protocols. *International Journal of Computational and Experimental Science and Engineering*, 10(3);420-426. <https://doi.org/10.22399/ijcesen.385>
- [54]güven, mesut. (2024). Dynamic Malware Analysis Using a Sandbox Environment, Network Traffic Logs, and Artificial Intelligence. *International Journal of Computational and Experimental Science and Engineering*, 10(3);480-490. <https://doi.org/10.22399/ijcesen.460>
- [55]Serap ÇATLI DİNÇ, AKMANSU, M., BORA, H., ÜÇGÜL, A., ÇETİN, B. E., ERPOLAT, P., ... ŞENTÜRK, E. (2024). Evaluation of a Clinical Acceptability of Deep Learning-Based Autocontouring: An Example of The Use of Artificial Intelligence in Prostate Radiotherapy. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1181-1186. <https://doi.org/10.22399/ijcesen.386>