SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I7P122

Volume 12 Issue 7, 280-294, July 2025

Original Article

A Hybrid Deep Learning Model with Generative Adversarial Network for Abnormality Detection from Surveillance Videos

Ganta Raju[1], G. S. Naveen Kumar[2]

[1]*Department of CSE, Mallareddy University, Hyderabad, Telangana, India.*
[2]*Department of Data Science & Information Technology Mallareddy University, Hyderabad, Telangana, India.*

[1]*Corresponding Author : rishi.shinny@gmail.com*

Received: 12 May 2025 Revised: 13 June 2025 Accepted: 14 July 2025 Published: 31 July 2025

Abstract *- Video anomaly detection is a real-world problem that artificial intelligence (AI) and computer vision applications can solve. In today's world, with our deteriorating environment, it is urgent to assess surveillance videos and process them in real-time for abnormalities to ensure the safety and security of citizens. Several deep-learning approaches have been developed to detect anomalies in videos. However, these traditional models require improvements in hybridization and utilize a Generative Adversarial Network (GAN) architecture to achieve enhanced performance. This paper presents a novel deep-learning framework that efficiently addresses this problem by searching surveillance footage for irregularities. A deep learning technique called GANDL-VAD is suggested for Video Abnormal Detection (VAD). It uses a GAN architecture and offers a hybrid DL model that considers both extracted and synthesized data, thereby improving the efficiency of detection and classification. The suggested hybrid deep learning model surpassed its modern rivals with an accuracy rate of 98.78%, according to experimental results on the UCF-Crime benchmark dataset. It can be used directly in current computer vision applications for video analytics and is capable of detecting anomaly events in surveillance videos.*

Keywords *- Video abnormality detection, Artificial Intelligence, Hybrid Deep Learning, Generative Adversarial Network, Computer vision.*

1. Introduction

Video surveillance is one of the most essential tools to keep the public safe and secure in the contemporary world. Going through surveillance footage from multiple cameras across a city manually is no small task. That is why you need to utilize Artificial Intelligence (AI) to analyze live surveillance videos and automatically identify unusual activities. In this way, any unusual activity can be recorded on video and then reported to the relevant authorities to enhance public safety and security [1]. Deep learning models are capable of processing video streams and detecting unusual events. One must realize that video analytics for abnormal event detection and alerting is a non-trivial task. A series of deep learning models, combined with preprocessing and feature engineering, can be deployed to detect suspicious actions in surveillance videos accurately. Analyzing different types of videos to detect anomalous activities is also significant in the case of social media [2].

For Computer Vision (CV) tasks, particularly video analytics on surveillance footage, deep learning models are often employed. Cloud computing infrastructure is also typical for storing and handling large amounts of data. Abnormality detection using deep autoencoders is a state-of-the-art approach that utilizes the detected frames as input to identify suspicious events within the video frames. In addition, deep learning models based on hybrid Neural Networks (NNs) that memorize models [3, 4] can efficiently recognize uncertain anomalies. It is also mentioned in the literature that multi-abnormality detection in videos can be enhanced by utilizing the Generative Adversarial Network (GAN) architecture to utilize training samples effectively. To push forward the cutting edge of video multi-abnormality detection, a comprehensive deep learning framework is to be developed using deep learning models over a GAN architecture.

However, current deep learning-based anomaly detection techniques still suffer from significant shortcomings. There are additional problems caused by the fact that most previous work focuses on spatial or temporal feature learning separately and only has partial feature representation. Others cannot generate additional training features for enhancing detection accuracy when only a small amount of labeled data is available. Additionally, current GAN-based video anomaly detection models fail to utilize the hybrid CNN and LSTM architectures fully. It leads to a bottleneck in detecting abnormal activities in complex surveillance environments when both appearance and motion cues are present. Thus, a joint GAN-based synthetic feature generation approach with reliable spatial and temporal feature learning in a hybrid deep learning framework is in great demand, which can lead to enhanced anomaly detection in surveillance videos.

Instead of separately processing spatial and temporal feature extraction for video anomaly detection, GANDL-VAD presents a joint, hybrid deep learning pipeline, as current methods do. Existing methods often condition on spatial content using CNNs for spatial feature learning and/or on temporal segmentation using LSTM for time decomposition; however, very few works have effectively combined these approaches for end-to-end training.

Furthermore, GANs have been applied to anomaly detection. Still, they are used only for reconstructing the input frames or producing anomaly scores, rather than synthesizing additional complementary feature presentations to enhance classification. The proposed model addresses these issues by integrating CNN-LSTM-based extracted features and synthetic features generated by a GAN architecture.

This combination of real and artificial features expands the feature space, thereby improving generalization, especially in scenarios with a small training set. In contrast to three cutting-edge models (HDLVAD, CNN+SRU, and Inter-fused Autoencoder), which separately study handcrafted features, temporal learning, or a hybrid model without GAN-based augmentation, the proposed method improves performance. According to experimental findings on the UCF-Crime benchmark, the GANDL-VAD approach outperforms the most advanced techniques in terms of precision, recall, and overall classification performance, achieving 98.78% anomaly detection accuracy compared to other methods.

Our study introduces a unique DL system that aims to accurately detect anomalies in surveillance data. A novel approach, specifically tailored to our problem, is developed: GANDL-VAD (Generative Adversarial Network-Based Hybrid Deep Learning for Video Abnormality Detection). This approach leverages the GAN architecture, hybridizing it with a deep learning model to improve the detection and classification of abnormal events by synthesizing and extracting features. Our hybrid deep learning model outperforms the competition in an empirical study based on the UCF-Crime dataset, achieving an accuracy of 98.78%.

This video anomaly detection deep learning framework can effectively be utilized in computer vision applications to detect abnormal events, particularly in surveillance videos. The rest of the paper is structured as follows: Section 2 analyzes deep learning algorithms for detecting and classifying abnormal events and video segments. Section 3 illustrates the proposed method for automatic anomaly detection in surveillance videos. The results of experiments with the UCF-Crime dataset are in Section 4. Lastly, Section 5 outlines directions for future research and issues worthy of further investigation.

2. Related Work

Numerous deep learning-based approaches exist for detecting video anomalies. Ilyas et al. [1] presented a hybrid deep learning technique that utilizes ResNet-101 models to identify crowd anomalies for temporal and spatial variables.

Manual motion descriptors were used to enhance detection accuracy on the PETS 2009 and UMN datasets. Garg et al. [2] focused on the secure transmission of social multimedia data and proposed a trust-based paradigm to enhance Quality Of Service (QOS) and dependability by utilizing Software-Defined Networking (SDN) and deep learning for anomaly detection. Zhou et al. [3] improved context and the decoder's extrapolation by introducing a hybrid autoencoder that enhances the LSTM encoder-decoder for video anomaly detection, outperforming current techniques in experiments.

Garg et al. [4] tested a hybrid approach to detecting anomalies in networks on benchmark datasets, which combined Improved-GWO and Improved-CNN, outperforming current approaches. Nayak et al. [5] focused on enhancing safety in public spaces through video surveillance systems, emphasizing the need for fast and precise anomaly detection. They pointed out that anomalies, such as fights or abandoned luggage, are often difficult to analyze due to ambiguity, varying settings, and a scarcity of facts.

Hamdi et al. [6] introduced a technique that effectively combines handcrafted features with a deep learning CNN to detect abnormalities. Post-processing increases accuracy by resolving false alarms. Rezaee et al. [7] emphasize the importance of automated crowd anomaly detection for public safety. Their paper investigates techniques for effective detection, such as deep learning. Ramchandran et al. [8] emphasize the importance of video anomaly detection for real-time applications in settings such as hospitals and airports. They present HDLVAD, an efficient deep-learning framework that outperforms current techniques.

Hussain et al. [9] report increased accuracy in anomaly detection by applying new sampling approaches that incorporate anomaly detection and machine learning. They constructed and evaluated hybrid models using video data. Deepak et al. [10] recommend utilizing cutting-edge multi-view representation learning techniques for anomaly identification in security footage. Hybrid and deep approaches fared better than innovative methods using benchmark datasets.

Jaouedi et al. [11] emphasized the importance of recognizing human actions in various applications. Their hybrid deep learning model, tested on the KTH dataset, achieved an impressive 96.3% accuracy in action recognition. On the other hand, Mansour et al. [12] combined Faster RCNN and DRL for detection and classification. Their IVADC-FDRL model attained accuracy rates of 98.50% and 94.80% on the UCSD dataset.

Aziz et al. [13] employed Mask-RCNN for appearance and unique intensity histogram-based motion descriptors, effectively integrating appearance and motion analysis.

Their connected component analysis yielded good accuracy on the Avenue and UMN datasets while reducing false alarms. Yu et al. [14] used motion and appearance cues, and their VEC technique improved video surveillance. A visual cloze is used to record high-level semantics; an exam was also created, and optical flow inference was employed. Lin et al. [15] presented a hybrid model that combined VAE for local characteristics and LSTM for long-term correlations to detect anomalies in time series without supervision, with results indicating that this approach outperformed others.

Aslam et al. [16] proposed the Inter-Fused Autoencoder (IFA), which integrates CNN and LSTM to identify video anomalies using deep learning. The IFA reconstructs with minimal error and efficiently learns spatiotemporal patterns. Various datasets demonstrate that our method outperforms Mean Squared Error (MSE), achieving AUC values of 96.005%, 93.0%, and 91.4%, respectively.

However, PSNR performed worse in comparison. Yang et al. [17] utilized a convolutional autoencoder network to analyze consecutive frames in a novel human-machine collaborative method for video anomaly identification. They produced a 93.7% AUC, surpassing the performance of the most advanced procedures. Computational efficiency and feedback inclusion will be the main focus areas in future studies.

Qasim et al. [18] proposed using a hybrid swarm intelligence method for identifying abnormal events in crowded films. The process builds a 2D variance plane and applies modified ACO and HOS techniques to identify motion abnormalities. The strategy surpassed the current approaches for the UMN and UCF datasets. A hybrid deep learning system, as suggested by Karadayi et al. [19], was designed specifically for COVID-19 trends to facilitate unsupervised anomaly detection in spatiotemporal data. The technique integrates spatial-temporal information and predicts epidemics early, outperforming existing models. Guha and Samanta [20] combined RPA and AI/ML to improve Anomaly Detection (AD) in title insurance. The hybrid autoencoder model addresses training and data variance issues while enhancing accuracy and business performance.

Li et al. [21] suggested a spatiotemporal model that combines a mixed Gaussian model for backdrop modeling with a CNN variant to extract features. The model utilizes Mahalanobis distance to provide precise location data, eliminating the need for specialized training to detect anomalous behavior. This model ensures accuracy, fast calculation, and high adaptability. Yang et al. [22] designed a model to identify unusual video activity by focusing on motionless and stationary behavior. Their framework employs a two-channel architecture with appearance and foreground characteristics. High-level feature learning and anomaly scoring are achieved using SDAE-DBN-PSVM architectures fused for event detection. According to the results, the model outperforms baseline methods on the MCG, UCSD, and Subway datasets.

Ghrib et al. [23] proposed an efficient hybrid LSTM Autoencoder-SVM model for identifying high-dimensional time series data abnormalities. Their strategy outperforms current techniques. Liu et al. [24] combined unsupervised and other models, expanding the scope of Video Anomaly Detection (VAD). This involves studying future trends, evaluating performance, and developing a taxonomy. Qasim et al. [25] proposed CNN + SRU models that achieve accuracies of 88.92% to 91.24% on UCF-Crime for video anomaly detection. Regarding accuracy, precision, and AUC, CNN + SRU outperforms other models. Integrating an attention layer will be a focus of future studies to enhance anomaly identification.

In their research, Shao et al. [26] emphasized that identifying anomalies is more successful than using baseline models. They suggested that future research should combine pixel-level object identification with collaborative training to identify anomalies in real-time. Similarly, Hassan et al. [27] focused on identifying network breaches in big data situations, proposing a hybrid deep CNN-WDLSTM IDS. On the UNSW-NB15 dataset, the WDLSTM achieved 97.1% accuracy by maintaining long-term relationships while extracting features using CNN.

Rahman et al. [28] achieved significant accuracy in identifying suicidal and depressed users, showcasing the efficacy of the DT-SVMNB model in practical situations. They also suggested that future research should focus on creating a real-time detection system and performing dynamic user interest analysis. Erhan et al. [29] addressed sensor system problems, discussing traditional and data-driven methodologies. They also highlighted the impact of architectures such as Cloud, Fog, and Edge on sensing techniques, pointing out open research questions and important sources for further study. Lastly, Santhosh et al. [30] looked into the use of visual surveillance to identify abnormalities in pedestrian behavior and traffic infractions, providing a detailed analysis and categorization system. They aimed to guide researchers on ITS matters.

Ullah et al. [31] used BD-LSTM and deep features in surveillance systems to create an intelligent anomaly detection framework. The framework showed increased accuracy on benchmark datasets. Jain et al. [32] presented a methodology for localizing and detecting gastrointestinal abnormalities. The model outperforms current techniques in terms of accuracy and segmentation. Protogerou et al. [33] examined the impact of the Internet of Things on security. They proposed using graph neural networks in a multi-agent system to identify anomalies in a distributed network architecture.

The approach uses few resources and is accurate. Elasayed et al. [34] suggested a hybrid OC-SVM and LSTM-autoencoder model for imbalanced datasets, which improves anomaly detection effectiveness. Tests on the InSDN dataset confirmed its efficacy, enhancing detection rates and significantly reducing processing time. Malik et al. [35] suggested a DL-driven approach for threat detection

based on CNN and LSTM. Testing revealed increased accuracy without sacrificing speed. They also mentioned plans to include more DL models and expand control plane capabilities.

Liu et al. [36] introduced a new annotation and assessment measure to improve anomaly detection. Their system, which focuses on anomalous regions, significantly enhances performance on standard datasets. Bozcan et al. [37] aimed to detect abnormalities for autonomous surveillance. Their UAV-AdNet system outperformed baseline methods in scene reconstruction and anomaly detection by utilizing GPS and image data.

Nasaruddin et al. [38] exploited robust background subtraction and attention regions in videos. Their algorithm achieved high accuracy using a substantial UCFCrime dataset, which aids security personnel in targeted investigations.

Hwang et al. [39] examined the minimum number of packet bytes required to detect anomalies in IoT traffic early, achieving high accuracy with few false positives using a combination of a CNN and an autoencoder. Doshi and Yilmaz [40] compared MONAD with other techniques and found that it provides rate limitations for asymptotic false alarms. MONAD uses statistical decision-making, and deep learning feature extraction is utilized for online video anomaly detection.

According to the literature reviewed, it is clear that while a few hybrid deep learning and GAN-based models have been developed for video anomaly detection, the majority of the models concentrate on handcrafted features or isolated spatial-temporal learning. Rare works have successfully integrated synthetic process feature generation with spatial and temporal process feature extraction under a unified scheme.

Furthermore, such an inductive algorithm identifies no comprehensive models that combine feature synthesis with extraction to enhance classification performance, particularly in cases involving small data samples.

To fill this gap, this paper proposes a hybrid deep learning architecture, GANDL-VAD, which integrates CNN-LSTM for feature extraction and GAN for generating synthetic features to enhance the performance of the anomaly detection task in surveillance videos.

3. Materials and Methods

This section outlines the proposed approach for automatically identifying and categorizing anomalies in surveillance footage. It uses a hybrid deep learning model and GAN architecture.

3.1. Problem Definition

The challenge lies in developing a deep learning-based framework that leverages a hybrid learning model and a GAN architecture to improve abnormal activity detection in surveillance videos.

3.2. Proposed Framework

In Figure 1, you can see how the different parts of our proposed architecture work together to achieve action recognition. First, our system utilizes transformation techniques to preprocess the raw data and compensate for the limited quantity of available datasets. Then, informative and discriminative features at the frame level will be extracted using a hybrid approach. Finally, an LSTM network processes the motion data to identify long-range dependencies and implicit correlations in the data sequence after the first identification round.

The process starts with a database of images, which is then expanded using the Augly library to create an augmented database. The augmented images undergo preprocessing steps, including resizing, normalization, and image extraction.

The processed photos are then sent into a hybrid DL model to extract characteristics. At the same time, a GAN architecture adapted from [33] is used to generate synthetic features. Both the extracted and synthetic features are then fed into a classification module consisting of a Fully Connected (FC) softmax layer, which assigns probabilities to different classes.

The result is a classification into one of several classes. This process combines data augmentation, deep learning, and advanced feature synthesis techniques to enhance classification performance. In the fourth stage, GAN is employed in the action recognition system to generate synthetic features.

Here, the GAN generates new synthetic features using the training dataset, which is larger due to the characteristics that the CNN-LSTM was able to obtain as inputs. Finally, in the fifth stage, an activation function for action classification based on softmax is fed the CNN-LSTM-retrieved features and the GAN-created synthetic features.

3.3. Data Augmentation

Techniques for enhancing data are increasingly common in various domains and are valuable for developing computer vision models. These strategies involve introducing variations to the input data to expand datasets without causing overfitting. Data augmentation also helps assess the robustness of trained models and enhance their overall performance. While image classification tasks can access a wide range of data augmentation approaches, video classification tasks have more limited options.

However, a few attempts have been made to enhance movies with data, albeit with fewer alternatives than image categorization. AugLy is an open-source data augmentation library known for augmentations, including superimposing films, combining movies, and simulating previously shared screenshots. Using AugLy, we applied four specific approaches to our model: adding noise, backdrops, blur, and adjusting brightness. Our model's training reliability and performance have been improved by artificially increasing the dataset and employing these techniques.

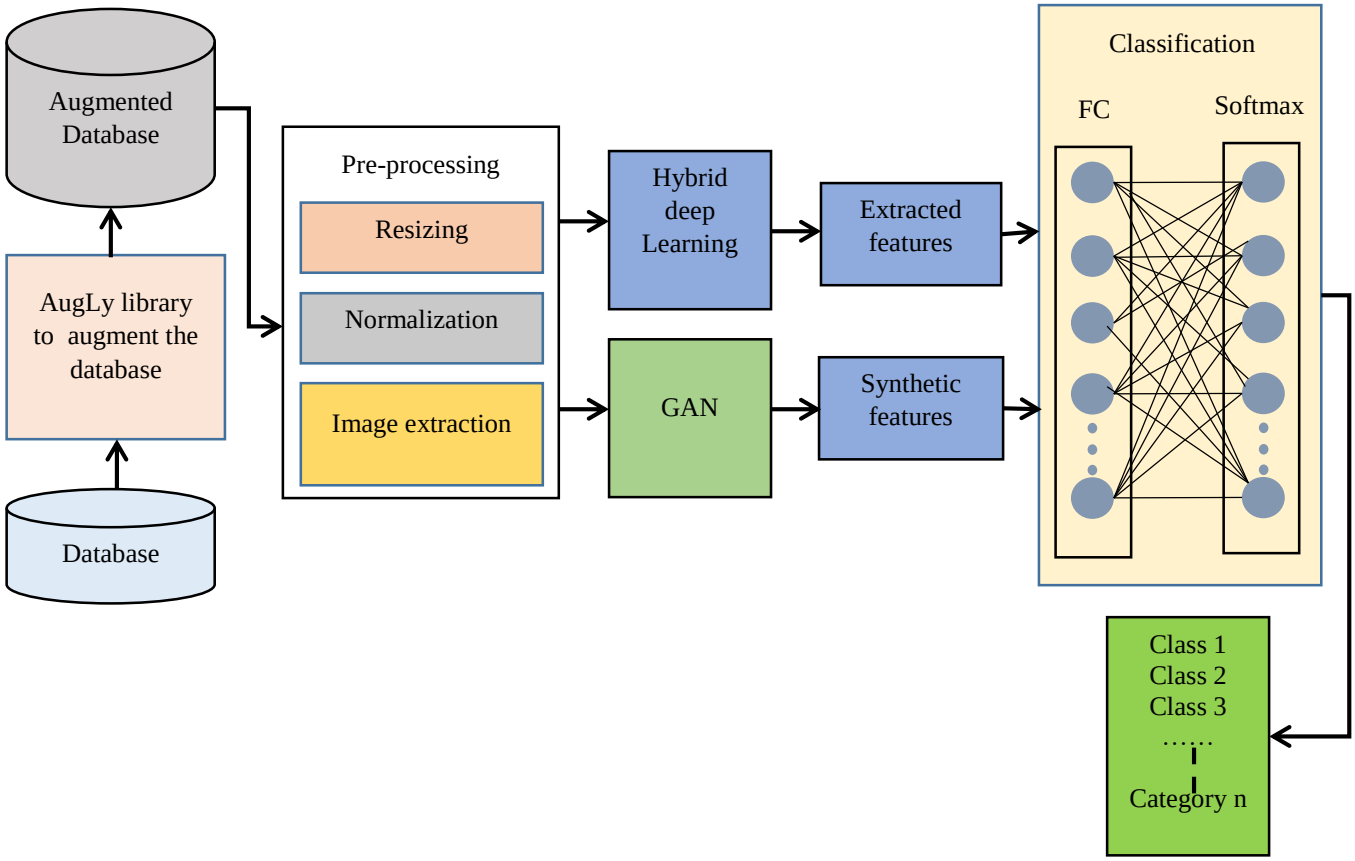

Fig. 1 Proposed GAN-based hybrid deep learning framework for video of normal multi-detection

3.4. Data Preprocessing

Depending on the needs of the machine learning model, various preprocessing methods can be employed when preparing data for machine learning. Here, preprocessing is used to detect abnormal human activity in videos. Three main steps are followed to prepare the information for further processing. First, removed the associated video frames and created a list of every picture in the movie, allowing us to manipulate the frames individually at a sample level. Then, resized the frames. Data preparation for machine learning can be accomplished using various preprocessing strategies, with the specific methods chosen depending on the data and the requirements of the machine learning model.

3.5. Feature Extraction Using a Hybrid Deep Learning Model

Using a hybrid architecture, our method for identifying irregularities in human behavior in films extracts both temporal and spatial characteristics from video frames. Video frames measuring 224x224x3 are sent into the model, which runs them through convolutional layers with progressively higher filter counts. Dropout layers, used to avoid overfitting, decrease the spatial dimensions of the feature maps by following each convolutional layer and a max pooling layer. After being processed, the feature maps are input into a sequence of LSTM layers to handle temporal dependencies between the video frames. Another dropout layer is used after the LSTM layers to reduce overfitting further. This combined architecture of CNNs and LSTMs for temporal sequence learning and spatial feature extraction is adequate for video analysis tasks, as it captures both spatial and temporal aspects. Figure 2 illustrates the hybrid neural network architecture for analyzing video frames and extracting features.

A 2D CNN-based framework is used to extract video spatial characteristics. CNNs, and specifically 2D-CNNs, are widely used in HAR research. Since constructed models are utilized to retrieve spatial information from every frame to interpret two-dimensional inputs, such as pictures, by modifying the input frame with various filters, a CNN can recognize different spatial characteristics, such as corners, edges, and textures. The geographical information in every video frame was extracted using the 2D-CNN in our investigation, and the 2D-CNN was applied to each of the twenty frames selected from each video. When evaluating the data, it is crucial to consider the film's spatial and temporal aspects. Our study had hardware constraints, including a GPU with specific RAM capacity. To work around these limitations, our network architecture was tailored to maximize the utilization of the available hardware assets. As part of our proposed paradigm, the CNN network architecture consists of a softmax output layer, four pooling layers, and four convolutional layers. Each 3 × 3 2D convolution kernel has a stride of 1 × 1.

During the first and second phases of maximum pooling, the kernel sizes are 4x4, while the rest of the max pooling layers are 2x2. The detailed architecture is explained in Figure 2.

Although CNNs help obtain spatial information from a single video frame, LSTMs are a superior option for obtaining time-related data.

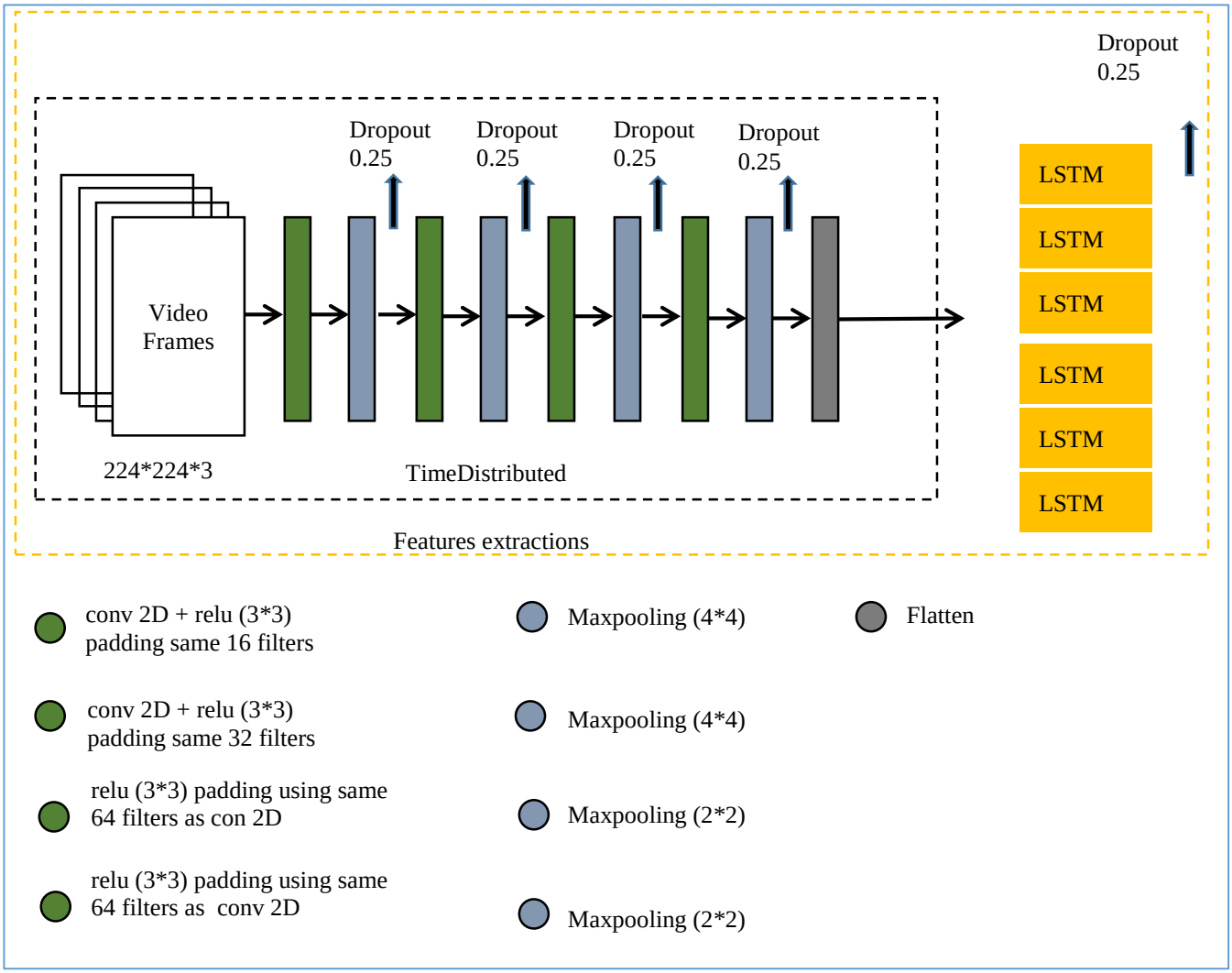

Fig. 2 Feature extraction using a hybrid deep learning model

LSTMs were created to obtain temporal dynamics from video frames. They are a type of RNN that handles sequential input, such as video, and are better suited for modeling short- and long-range sequence feature relationships than conventional RNNs. The three basic "gates" that comprise an LSTM unit are the input, output, and forget gates [28]. These gates control information transit between previous temporal stages and hidden states. The LSTM network can continue to store information about previous inputs internally and utilize it to forecast the following inputs.

An LSTM network can analyze a sequence and identify the temporal connections between video frames for human action recognition. In our implementation, we employed thirty-two components in an LSTM network. The LSTM network obtains spatial features from the CNN input for temporal feature extraction. Using the geographical data that CNN has gathered and prior inputs, the LSTM network can predict temporal correlations in the video.

In conclusion, LSTMs are more suitable for capturing temporal information, whereas CNNs excel at extracting spatial information from individual video frames. Extraction of temporal and spatial information can be done effectively from video data by combining CNN and LSTM.

The hybrid CNN-LSTM approach's intricacy at each step in time can be computed by adding the complexities of the CNN and LSTM layers, as shown below: While the exact computation is made for every training process, $O\left(\sum_{l=1}^{d}(n_{l-1}.s_l^2.n_l.m_l^2)+w\right)$,
$O\left(\left(\sum_{l=1}^{d}(n_{l-1}.s_l^2.n_l.m_l^2)+w\right).i.e\right)$.

In this case, the numbers d, n_l, epochs, i, and w represent how many convolutional layers there are, sl, ml, the input channel count, LSTM weights, input length, and input epochs, and n_(l-1), the generated feature map's spatial dimension. Using the conventional asymptotic notation, can say that our model has complexity O.

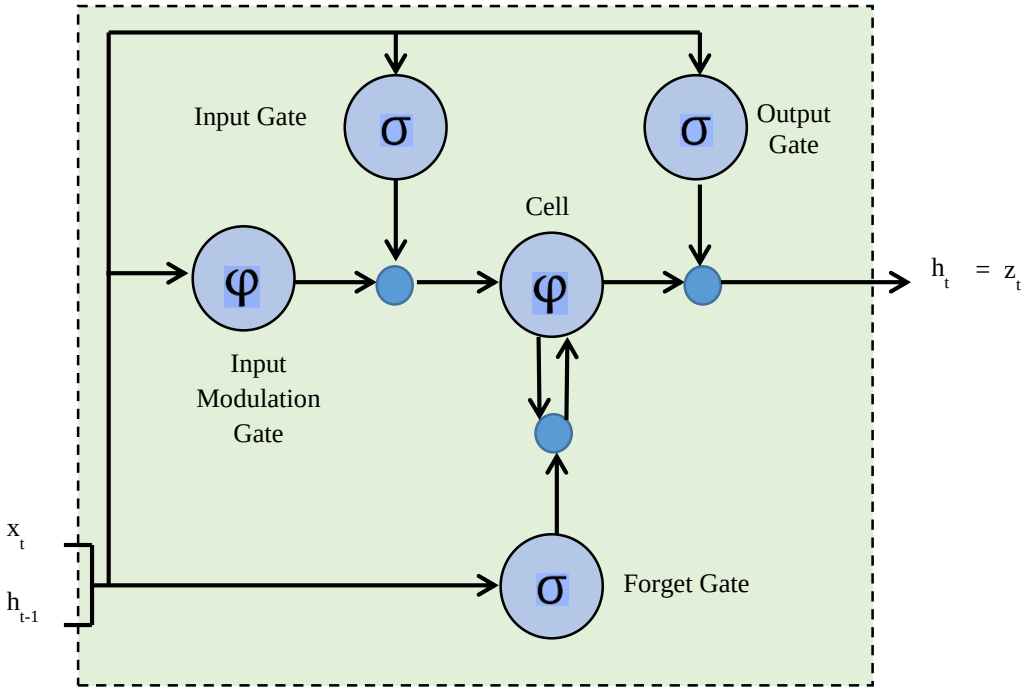

Fig. 3 Architectural overview of the LSTM unit

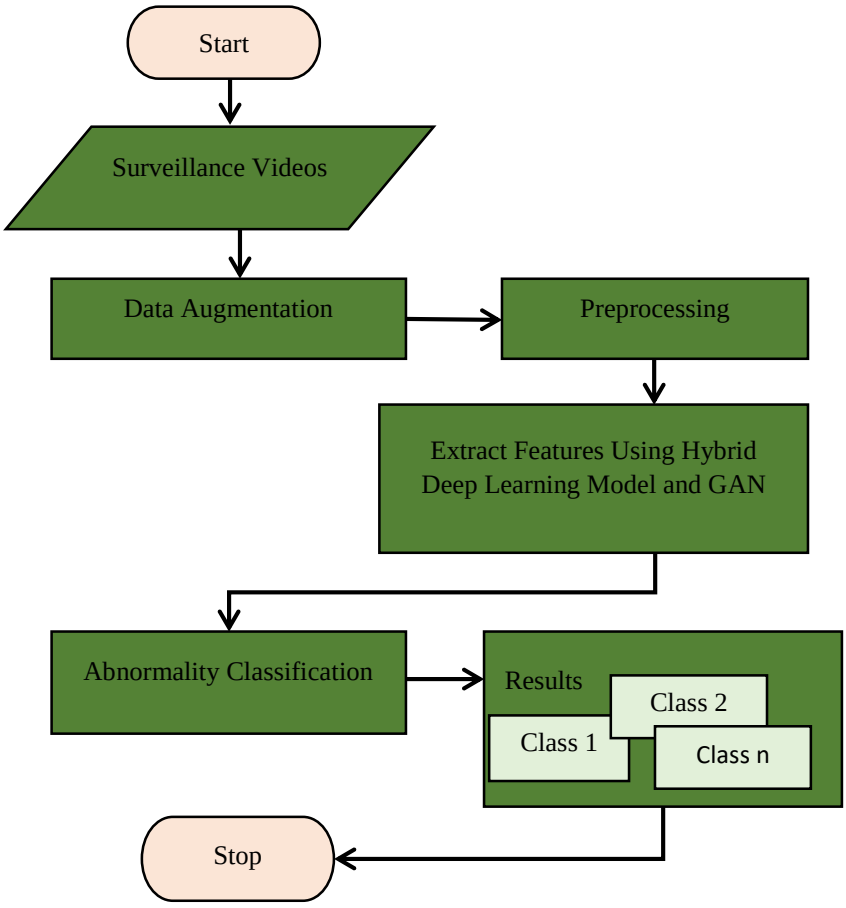

Fig. 4 Flowchart for the suggested system operation

Figure 4 shows how a proposed system analyzes surveillance videos. It starts with collecting surveillance videos, which are then enhanced through data augmentation. Next, the information is preprocessed to get it ready for feature extraction. The system uses a combination of DL and Generative Adversarial Networks (GAN) to extract relevant features from the preprocessed data. These features are then used for abnormality classification, where the model groups the data into different classes (e.g., Class 1, Class 2, up to Class n) based on the detected abnormalities. The results of this classification give insights into various abnormal activities in the surveillance footage. The process ends when the system operation is terminated. This flowchart outlines the steps involved in the system, from data collection to the final classification output in a clear and structured manner.

3.6. GAN for Synthetic Features

The initial phase trains a GAN model using spatiotemporal data to generate synthetic features for human activity detection. The design of the GAN model is illustrated in Figure 4. First, a capsule encodes the retrieved features instead of enriching the video visuals directly. By using this approach, artificial data is generated to recognize human activities that are likely to be less challenging to process. Avoiding the processing and storage of more photos may not be feasible. The Architecture representation and training of GAN are shown in Figure 5. It starts with a noise input into the Generator for synthetic features. The synthetic features are then passed to the Discriminator, which determines whether the product is authentic or fake. The Discriminator also evaluates real features. The Discriminator is the simplest neural network in this context, and it is trained using a loss function that determines the Discriminator's loss based on its ability to classify natural and artificial features. Use this loss to update the Discriminator. At the same time, the performance of the Generator is evaluated using the same loss function to minimize the loss, thereby forcing the Generator to create more realistic synthetic features. The Generator loss and Discriminator loss are used in a feedback loop to iteratively refine both models iteratively, increasing the Discriminator's capacity to differentiate between real and fake features, and, in essence, also growing the Generator's capacity to generate realistic features.

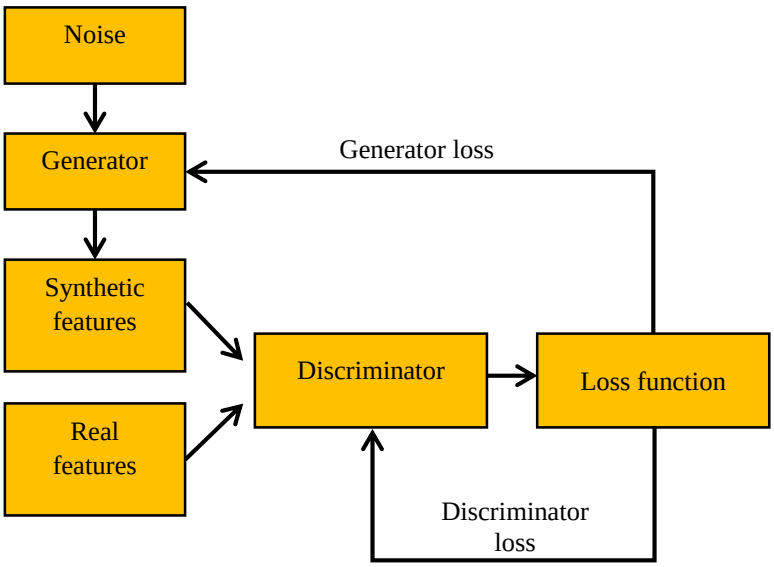

Fig. 5 Overview of the GAN architecture adapted from [33]

To understand the GAN-GP [33], it is essential to comprehend how a GAN works [32]. Adversarial networks are a relatively new term introduced by Goodfellow as a generic term for a GAN (Generative Adversarial Network). They consist of a generator neural network and a discriminator that collaborate to create synthetic data. In this sense, the Discriminator distinguishes whether the artificial data is real or not, as it compares it against the original one, while the Generator generates fake data with some noise. Train the Discriminator and Generator to improve their ability to distinguish between real and fake data. This type of power plant has been trained to provide more realistic-looking synthetic data. When playing a minimax game, these two networks are trained adversarially; both players are jointly learnt. Creating synthetic data that is as realistic as possible, the Generator's objective is to transform samples into real input data from a simple noise distribution (such as a Gaussian or uniform distribution). However, it is trained to distinguish between authentic and fraudulent input data. Use the minimax game training method, which helps generate data with a distribution as close as possible to the actual distribution. The GAN initially suggested had convergence and stability problems, making it difficult to assess the quality of the tuning settings. The only way users had to check whether the GAN model was well-trained was through the generated samples.

The GAN in [33] is employed in this study to extract artificial characteristics. When a GAN uses the Wasserstein distance (W (P, Q)), it calculates the distance between

locations in the P and Q distributions. The Wasserstein distance replaces the GAN value function, which exhibits Jensen-Shannon divergence. Many problems with the first GAN algorithm have been resolved by the Wasserstein GAN method, which is also used to generate realistic data. However, GAN has drawbacks, including stability problems and a slow training rate. A proposal for GAN-GP was made in [33] to address these issues. GAN-GP confines the discriminator neural network using a gradient penalty to provide regular gradients, which is the primary distinction between GAN and GAN-GP. This enables faster and more stable convergence while preventing training oscillations. The norm of the discriminator output gradient is used as a cost function to apply the gradient penalty. Due to their ability to generate realistic samples and understand complicated data distributions, generative adversarial networks, or GANs, were originally selected for feature synthesis. Gradient-penalized Wasserstein GAN (GAN-GP) was specifically chosen due to its demonstrated superior stability and training speed compared to the original GAN technique. In the given example, G(z,c), the Generator uses a condition class label c and random noise z_{p} to construct a feature map. Using optimization, the discriminant D (f,c) generates a value given the inputs f and a condition class c.

$$E_{f\sim P_r}[D(f)] - E_{\hat{f}\sim P_g}\left[D\left(\hat{f}\right)\right] + \lambda E_{\hat{f}\sim P_{\hat{f}}}\left[\left(\left\|\nabla_{\hat{f}} D\left(\hat{f}\right)\right\|_2 - 1\right)^2\right]$$

The generator distribution is denoted by P_g and the real input feature distribution by P_{r}.

3.7. Abnormal Activity Recognition

One of the key components of our action categorization process is the classification of actions. This can be achieved by GAN-based feature synthesis and traits obtained by the CNN-LSTM. The dataset's class space dictates the output size of the last dense layer for classification. The Softmax function [33] is in this final thick layer. A well-liked activation function for multi-class classification issues is softmax. After determining the probability of each class for the given sample, it identifies the group most likely to occur. This capability simplifies understanding the classification results by providing probabilities for each class.

3.8. Proposed Algorithm

GANDL-VAD: Generative Adversarial Network-based Hybrid Deep Learning for Video Abnormal Detection. This method utilizes a hybrid DL technique based on a Convolutional Generative GAN and other deep learning methods. The objective of the algorithm is to enhance the precision and robustness of abnormal behavior detection in videos by training a classifier using features derived from both real and artificial images generated by the GAN. The classifier helps distinguish between normal and abnormal events in the video clip.

Algorithm 1: Generative Adversarial Network-Based Hybrid Deep Learning for Video Abnormal Detection (GANDL-VAD)

Algorithm: Generative Adversarial Network-Based Hybrid Deep Learning for Video Abnormal Detection (GANDL-VAD)
Input: dataset D for UCF-Crime
Output: Video anomaly detection results R, performance indicators P

1. Begin
2. D'←AugmentData(D)
3. D'←Preprocess(D)
4. (T1, T2)←SplitData(D')
5. Configure hybrid DL model m1 (as in Figure 2)
6. Compile m1
7. Configure GAN model m2 (as in Figure 4)
8. Compile m2
9. m1'←TrainHybridDL(T1)
10. realFeatures←TestHybridDL(m1', T2)
11. m2'←TrainGAN(T1)
12. syntheticFeatures←TestGAN(realFeatures, noise)
13. m←TrainClassifier(realFeatures, syntheticFeatures) (as in Figure 1)
14. R←DetectAbnormalities(T2, m)
15. P←PerformanceEvaluation(R, ground truth)
16. Print R
17. Print P
18. End

Algorithm 1 utilizes a Deep Learning Method based on GANs to detect abnormal activities in videos. The UCF-Crime dataset is required as input, and it outputs the video abnormality detection results and performance statistics. It begins with the dataset, which is processed and refined into a trainable dataset. Then, the data is split into two sets: T1 and T2, corresponding to the training and test sets. On subset T1, train a hybrid deep learning model (m1). At the same time, a GAN model (m2) is trained on portion T1 to generate synthetic features. Finally, a classifier is trained using the natural features obtained from the hybrid model and the synthetic features generated by the GAN to detect

abnormalities in subset T2. The detected results are then compared with the data by the algorithm to give some performance statistics. Finally, the detected anomalies and statistics are printed to evaluate the performance and propose an algorithm that leverages both synthetic and natural features to improve video anomaly detection.

3.9. Dataset Details

UCF-Crime Dataset: This work utilized the UCF-Crime dataset [41], which contains video clips labeled with various types of crimes. This method is widely used in studies involving video surveillance and activity recognition. The dataset includes videos of multiple crimes, including robbery, fighting, and burglary. This dataset is used to develop and evaluate algorithms for video-based crime detection.

3.10. Evaluation Methodology

Especially for video anomaly detection, the confusion matrix helps evaluate the performance of classification algorithms. In videos, the True Positive (TP) indicates that the model predicts an abnormality, and the True Negative (TN) suggests that the model predicts normality. However, if the model mispredicts an abnormality, it performs a False Positive (FP), which is a Type I error. If it mispredicts normality, it makes a Type II error, or a False Negative (FN).

Precision (p) = TP / (TP + FP)

Recall (r) = TP / (TP + FN)

F1-score = 2 * (precision * recall) / (precision + recall)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Through the examination of the confusion matrix, metrics such as F1 score, recall, accuracy, and precision can be calculated. These metrics offer insight into the model's overall predictive ability and ability to identify video anomalies.

4. Experimental Results

The following section presents the experimental results for the automatic detection and categorization of abnormal events, such as nakedness, in the test films under the proposed framework. Some cutting-edge DL models, including MobileNet, UNet, and VGG-19, are compared to the outcomes of the suggested deep learning model. The MobileNet is a lightweight deep-learning model for mobile and embedded devices. Developed by Google, it is known for its efficient memory and processing power use. One of the more frequently used CNN architectures is UNet. The proposal was made by researchers at the University of Freiburg in 2015. The architecture in UNet is distinctive and well-known due to its U-shape, which combines upsampling and downsampling to capture context while providing localization information. Belonging to the VGG series of models, VGG-19 is a CNN architecture. Sixteen convolutional layers and three FC layers make up the exact 19 layers of VGG-19. Small 3x3 convolution filters throughout the network contribute to the architecture's renowned homogeneity and simplicity. VGG-19 has been extensively utilized in computer vision applications due to its simplicity of design and efficacy, particularly in object detection and image categorization. Our experimental setup includes Anaconda, the Python data science platform, and Python 3.12. The empirical experiment was conducted on a Windows 11 computer equipped with a 13th Gen Intel(R) Core(TM) i7-1355U CPU, operating at 1.70 GHz, featuring 10 cores, 12 logical processors, and 16 GB of RAM.

Fig. 6 An excerpt from labeled data

Figure 6 presents an excerpt from the dataset with two different video frames reflecting "fights" and "no fights" labels.

Fig. 7 Experimental results about video abnormal detection

As presented in Figure 7, the results of detecting abnormalities in the given video for different frames are provided.

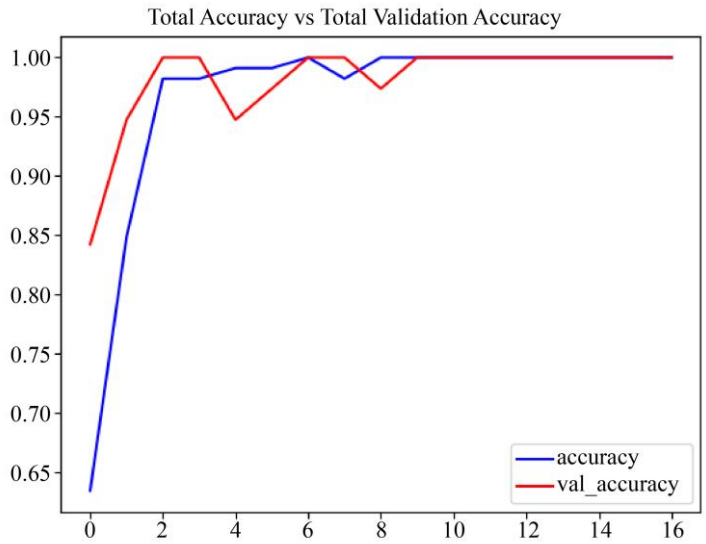

Fig. 8 Accuracy of the proposed deep learning model

A machine learning model's overall accuracy and validation accuracy during 16 training epochs are displayed in Figure 8. The accuracy values on the y-axis range from 0.65 to 1.00, whereas the x-axis has a range of 0 to 16 epochs. While the red line represents validation accuracy, the blue line represents training accuracy. The two accuracies increase sharply initially, with the training accuracy reaching more than 98% around the 4th epoch. The validation accuracy closely follows this trend, nearing more than 98% but with slight fluctuations. After the initial rise, both accuracies remain consistently high. The training accuracy shows minor variability, while the validation accuracy maintains a more stable trend. This indicates low overfitting and good model performance on the training and validation datasets.

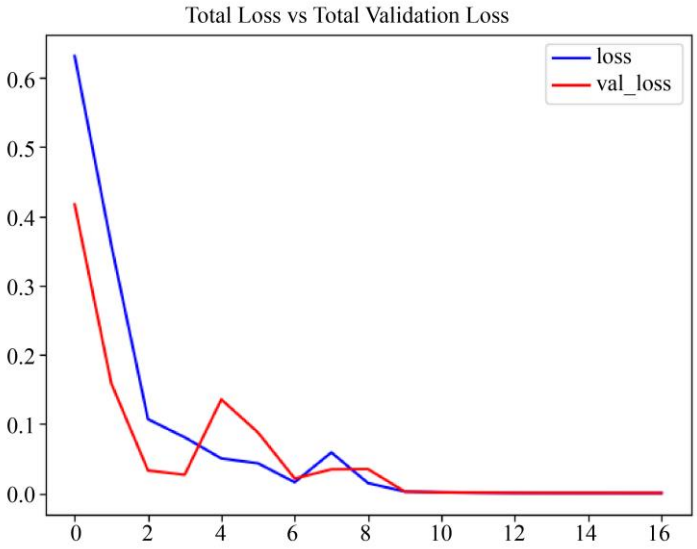

Fig. 9 Loss dynamics of the proposed hybrid learning model

The total loss and total validation loss over 16 epochs of machine learning model training are compared in Figure 9. The x-axis shows the number of epochs, which goes from 0 to 16, while the y-axis shows the loss values, which vary from 0 to 0.6. The red line represents the validation loss, while the blue line indicates the training loss. Both lines first barely decline slightly. The training loss starts at approximately 0.6 and drops to almost zero by the fourth epoch. Comparable in pattern, the validation loss begins at a lower value than the training loss and, with some variations, approaches zero by the fourth epoch. Both losses remain close to zero for the remaining epochs, indicating consistent model functionality improvement. This suggests a slight overfitting, but the model is adequately trained.

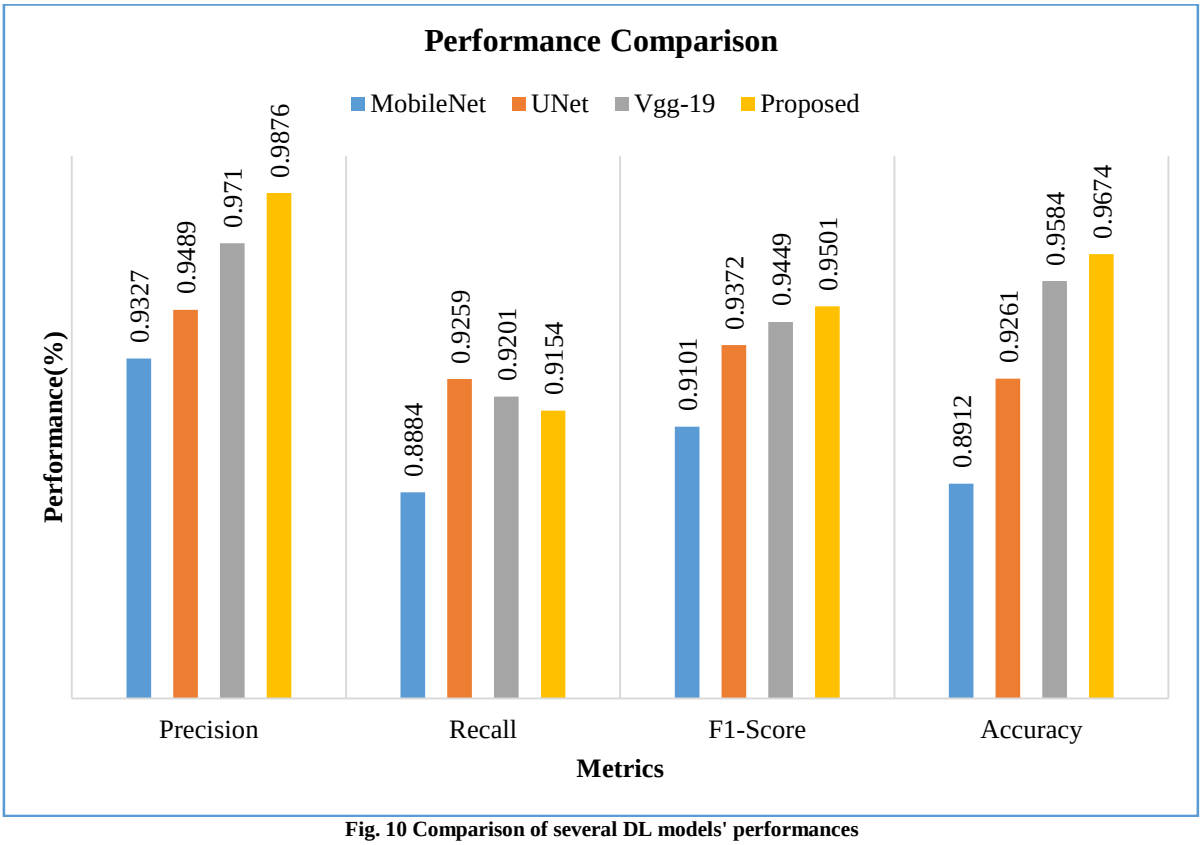

Fig. 10 Comparison of several DL models' performances

Accuracy, F1-Score, Precision, and Recall are the four measures used to compare the performance of four distinct models (MobileNet, UNet, Vgg-19, and the suggested hybrid model) in Figure 10. With an accuracy of 0.9674, an F1-score of 0.9501, a recall of 0.954, and a precision of 0.9876, the proposed model outperforms the previous models in each of the four criteria. VGG-19 also shows strong performance, particularly in precision (0.971) and F1-Score (0.9499), closely following the suggested model. UNet performs well in all metrics but does not exceed Vgg-

19 or the proposed model. In contrast, MobileNet consistently ranks the lowest, with precision 0.9327, recall 0.8884, F1-Score 0.9101, and accuracy 0.8932 among the four models. After comparing the suggested hybrid deep learning model to all other deep learning models, it may be concluded that it performs the best.

5. Discussion

The development of artificial intelligence has made computer vision applications increasingly crucial for solving problems in various domains. In urban areas, people encounter different incidents, including abnormal activities. To ensure public safety and security, authorities have been utilizing video surveillance to detect irregular behavior and gather evidence in the event of untoward incidents. Deep learning models, particularly those employing supervised learning, are highly valuable for analyzing video data and detecting abnormalities. In this paper, a GAN-based feature extraction method is proposed, utilizing synthetic feature extraction to enhance the performance of detecting abnormal events. Additionally, a hybrid model is proposed to process spatial and temporal features. The framework and models described in this paper could enable the automatic detection of abnormal events in surveillance videos. From the reviewed literature, it is evident that there are only a few hybrid deep learning and GAN-based approaches developed in the context of video anomaly detection, as most existing models tend to focus on handcrafted features or separate spatial and temporal learning. Only a few works have effectively combined synthetic process feature generation with spatial and temporal process feature extraction within a common framework. Moreover, this inductive algorithm does not identify any existing general models that perform joint feature synthesis and extraction to enhance classification accuracy, particularly for small training samples. To address this gap, we propose a type of hybrid deep learning model, namely GANDL-VAD in this paper, which combines CNN-LSTM for extracting features from observed data and GAN for generating synthetic features to enhance anomaly detection performance in surveillance videos. However, as section 5.1 explains, the system has several drawbacks.

5.1. Limitations

Some limitations of the system proposed in this paper are discussed. First, this dataset has a relatively small sample size, so these findings are not generalizable without further examination of a wider array of samples. Second, there is a need for improvements in the loss function related to the GAN architecture to minimize the error rate. Improve the suggested framework for recognizing crowd behavior in surveillance videos.

6. Conclusion and Future Work

Developed a comprehensive framework that combines multiple DL methods for the specific task of anomaly detection in surveillance videos. GANDL-VAD is a novel approach for video abnormal detection based on the GAN architecture and a hybrid DL model. This model utilizes a synthetic and extracted feature model to enhance the detection and classification efficiency. In video feature extraction, GANs automatically learn from video data and extract features relevant to the task at hand. More specifically, the generator network produces fake video samples, and the discriminator network evaluates the generated videos to identify whether each one is real or fake. The other player (the Generator) improves at producing realistic video samples, while the Discriminator becomes more adept at differentiating between actual and fake movies through such an adversarial process. In an empirical study (to test the model on a benchmark dataset called UCF-Crime), our hybrid deep-learning model outperformed the previous model with an accuracy of approximately 98.78%. This deep learning framework can be applied to existing computer vision applications to detect atypical behavior in security videos. It can be planned to enhance our framework for preventing, detecting, and classifying crowd behavior in surveillance videos in the future.

References

[1] Zirgham Ilyas et al., "A Hybrid Deep Network Based Approach for Crowd Anomaly Detection," *Multimedia Tools and Applications*, vol. 80, pp. 24053-24067, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sahil Garg et al., "Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 566-578, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fuqiang Zhou et al., "Unsupervised Learning Approach for Abnormal Event Detection in Surveillance Video by Hybrid Autoencoder," *Neural Processing Letters*, vol. 52, pp. 961-975, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sahil Garg et al., "A Hybrid Deep Learning-Based Model for Anomaly Detection in Cloud Datacenter Networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 924-935, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das, "A Comprehensive Review on Deep Learning-Based Methods for Video Anomaly Detection," *Image and Vision Computing*, vol. 106, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Slim Hamdi et al., "Hybrid Deep Learning and HOF for Anomaly Detection," *2019 6th International Conference on Control, Decision and Information Technologies*, Paris, France, pp. 575-580, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Khosro Rezaee et al., "A Survey on Deep Learning-Based Real-Time Crowd Anomaly Detection for Secure Distributed Video Surveillance," *Personal and Ubiquitous Computing*, vol. 28, pp. 135-151, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Anitha Ramchandran, and Arun Kumar Sangaiah, "Unsupervised Deep Learning System for Local Anomaly Event Detection in Crowded Scenes," *Multimedia Tools and Applications*, vol. 79, pp. 35275-35295, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Fizza Hussain et al., "Revisiting the Hybrid Approach of Anomaly Detection and Extreme Value Theory for Estimating Pedestrian Crashes Using Traffic Conflicts Obtained from Artificial Intelligence-Based Video Analytics," *Accident Analysis & Prevention*, vol. 199, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] K. Deepak et al., "Deep Multi-view Representation Learning for Video Anomaly Detection Using Spatiotemporal Autoencoders," *Circuits, Systems, and Signal Processing*, vol. 40, pp. 1333-1349, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Neziha Jaouedi, Noureddine Boujnah, and Med Salim Bouhlel, "A New Hybrid Deep Learning Model for Human Action Recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447-453, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Romany F. Mansour et al., "Intelligent Video Anomaly Detection and Classification Using Faster RCNN with Deep Reinforcement Learning Model," *Image and Vision Computing*, vol. 112, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Zafar Aziz et al., "Video Anomaly Detection and Localization Based on Appearance and Motion Models," *Multimedia Tools and Applications*, vol. 80, pp. 25875-25895, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Guang Yu et al., "Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events," *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA USA, pp. 583-591, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shuyu Lin et al., "Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 4322-4326, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nazia Aslam, and Maheshkumar H. Kolekar, "Unsupervised Anomalous Event Detection in Videos Using Spatio-Temporal Inter-Fused Autoencoder," *Multimedia Tools and Applications*, vol. 81, pp. 42457-42482, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Fan Yang et al., "Human-Machine Cooperative Video Anomaly Detection," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Tehreem Qasim, and Naeem Bhatti, "A Hybrid Swarm Intelligence Based Approach for Abnormal Event Detection in Crowded Environments," *Pattern Recognition Letters*, vol. 128, pp. 220-225, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yildiz Karadayi, Mehmet N. Aydin, and Arif Selçuk Öğrencí, "Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data Using Deep Learning: Early Detection of COVID-19 Outbreak in Italy," *IEEE Access*, vol. 8, pp. 164155-164177, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Abhijit Guha, and Debabrata Samanta, "Hybrid Approach to Document Anomaly Detection: An Application to Facilitate RPA in Title Insurance," *International Journal of Automation and Computing*, vol. 18, pp. 55-72, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Zhaoyan Li, Yaoshun Li, and Zhisheng Gao, "Spatiotemporal Representation Learning for Video Anomaly Detection," *IEEE Access*, vol. 8, pp. 25531-25542, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Meng Yang et al., "Deep Learning and One-Class SVM Based Anomalous Crowd Detection," *2019 International Joint Conference on Neural Networks*, Budapest, Hungary, pp. 1-8, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Zeineb Ghrib, Rakia Jaziri, and Rim Romdhane, "Hybrid approach for Anomaly Detection in Time Series Data," *2020 International Joint Conference on Neural Networks*, Glasgow, UK, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yang Liu et al., "Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1-38, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Maryam Qasim, and Elena Verdu, "Video Anomaly Detection System Using Deep Convolutional and Recurrent Models," *Results in Engineering*, vol. 18, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Wenhao Shao et al., "COVAD: Content-Oriented Video Anomaly Detection Using a Self Attention-Based Deep Learning Model," *Virtual Reality & Intelligent Hardware*, vol. 5, no. 1, pp. 24-41, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Mohammad Mehedi Hassan et al., "A Hybrid Deep Learning Model for Efficient Intrusion Detection in Big Data Environment," *Information Sciences*, vol. 513, pp. 386-396, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Md. Shafiur Rahman et al., "An Efficient Hybrid System for Anomaly Detection in Social Networks," *Cybersecurity*, vol. 4, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] L. Erhan et al., "Smart Anomaly Detection in Sensor Systems: A Multi-Perspective Review," *Information Fusion*, vol. 67, pp. 64-79, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] K.K. Santhosh, Debi Prosad Dogra, and Partha Pratim Roy, "Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1-26, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Waseem Ullah et al., "CNN Features with Bi-Directional LSTM for Real-Time Anomaly Detection in Surveillance Networks," *Multimedia Tools and Applications*, vol. 80, pp. 16979-16995, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[32] Samir Jain et al., "A Deep CNN Model for Anomaly Detection and Localization in Wireless Capsule Endoscopy Images," *Computers in Biology and Medicine*, vol. 137, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Aikaterini Protogerou et al., "A Graph Neural Network Method for Distributed Anomaly Detection in IoT," *Evolving Systems*, vol. 12, pp. 19-36, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[34] Mahmoud Said Elsayed et al., "Network Anomaly Detection Using LSTM Based Autoencoder," *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, Alicante Spain, pp. 37-45, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[35] Jahanzaib Malik et al., "Hybrid Deep Learning: An Efficient Reconnaissance and Surveillance Detection Mechanism in SDN," *IEEE Access*, vol. 8, pp. 134695-134706, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[36] Kun Liu, and Huadong Ma, "Exploring Background-Bias for Anomaly Detection in Surveillance Videos," *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France, pp. 1490-1499, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[37] Ilker Bozcan, and Erdal Kayacan, "UAV-AdNet: Unsupervised Anomaly Detection using Deep Neural Networks for Aerial Surveillance," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Las Vegas, NV, USA, pp. 1158-1164, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Nasaruddin Nasaruddin et al., "Deep Anomaly Detection through Visual Attention in Surveillance Videos," *Journal of Big Data*, vol. 7, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[39] Ren-Hung Hwang et al., "An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection," *IEEE Access*, vol. 8, pp. 30387-30399, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[40] Keval Doshi, and Yasin Yilmaz, "Online Anomaly Detection in Surveillance Videos with Asymptotic Bound on False Alarm Rate," *Pattern Recognition*, vol. 114, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[41] Real-world Anomaly Detection in Surveillance Videos, UCF, 2016. [Online]. Available: https://www.crcv.ucf.edu/projects/real-world/