Journal of Theoretical and Applied Information Technology

30th September 2023. Vol.101. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TWO STREAM SPATIAL-TEMPORAL FEATURE EXTRACTION AND CLASSIFICATION MODEL FOR ANOMALY EVENT DETECTION USING HYBRID DEEP LEARNING ARCHITECTURES

[1.] P. MANGAI,, [2.] M. KALAISELVI GEETHA, [3.] G. KUMARAVELAN
Research Scholar,
Department of Computer Science & Engineering,
Annamalai University, Annamalainagar, Tamilnadu, India.
Professor,
Department of Computer Science & Engineering,
Annamalai University, Annamalainagar, Tamilnadu, India.
Assistant Professor,
Department of Computer Science,
Pondicherry University, Karaikal, Puducherry, India,
E-mail: pannirselvammangai@gmail.com, geesiv@gmail.com, gkumaravelanpu@gmail.com

ABSTRACT

Identifying events using surveillance videos is a major source that reduces crimes and illegal activities. Specifically, abnormal event detection gains more attention so that immediate responses can be provided. Video processing using conventional techniques identifies the events but fails to categorize them. Recently deep learning-based video processing applications provide excellent performances however the architecture considers either spatial or temporal features for event detection. To enhance the detection rate and classification accuracy in abnormal event detection from video keyframes, it is essential to consider both spatial and temporal features. Based on this, two-stream hybrid deep learning architectures like YOLOV4 with VGG16, Optical FlowNet with VGG16, and CNN-LSTM has been presented in this research work. The two-stream architecture handles the spatial features using hybrid YOLOV4, temporal features are handled using hybrid Optical FlowNet and finally, the features are concatenated and classified using CNN-LSTM based deep learning architecture. The proposed model attains maximum accuracy of 95.6% which indicates better performance compared to state of art of techniques.

Keywords: *Anomaly Event Detection, Keyframe Extraction, Spatial And Temporal Feature, Deep Learning, YOLOV4, Optical Flownet, VGG-16, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM).*

1. INTRODUCTION

Abnormal event detection becomes much more familiar due to the effective utilization of surveillance applications in the home, public places, offices, and industries. However, the conventional video processing methodologies cannot able to detect the abnormalities as a real-time application. It is essential to automate the abnormal event detection so that time and cost can be minimized and suitable actions can be taken before any big issues. Conventional anomaly event detection techniques like the histogram of oriented gradients, and histogram of oriented flows use hand-crafted features to process the video. However, these methodologies are not robust and their accuracy is very low for complex scenarios. With the major objective of improving the detection accuracy in abnormal event detection multiple combinations of deep learning models are presented in this research work.

The basic principle behind abnormal event detection is categorized into object-based and frame-based processes. In the object-based event detection, objects are detected in the input frames and an anomaly score has been generated for each object. Based on the object anomaly score it infers the anomaly of a frame. This object-based detection is simple and less affected by the noise as it uses detected objects in the frame. The detection accuracy of the object-based approach is high even while handling multiple scenes. The major limitation of object-based detection is its initial preprocessing which increases the

computation time. Also, the detection performance is based on the object and if the objects are not detected then this method will produce false results in the anomaly event detection. Object-based detection has inconsistent inference time as it infers the anomaly score of all detected objects in the frame which slow down the detection process. Due to this reason object-based anomaly detection is not preferred for real-time detection applications.

Vision-based recognition for surveillance applications utilizes motion projection profile features to extract the keyframes from the videos. The extracted key frames are further clustered using k-means, Gaussian mixture model, etc., to detect the anomaly [1]. Another popular method of anomaly detection is frame based approach in which frames are used as input to the model. Instead of generating an anomaly score for objects, it generates an anomaly score for frames. Unlike object-based detection, frame-based detection neglects the initial preprocessing which makes the approach suitable for real-time applications. The interference time of frame-based detection is consistent which doesn't slow down the detection process. Machine learning-based anomaly event detection is evolved in the past years. The features obtained from the video frames are classified using machine learning algorithms to define the anomaly status. However, the performances of machine learning-based approaches are not satisfactory as it provides limited results. Moreover, these models require separate feature processing units before classification which increases the overall computation cost. Recently deep learning has proven the maximum ability in various image processing applications [2-3]. The performances of deep learning models are much better than traditional machine learning approaches like support vector machines, random forests, decision trees, etc., for better performance hybrid deep learning techniques are used in recent applications. Based on these observations, the proposed architecture is composed using different deep learning techniques to attain better detection accuracy in the anomaly event detection from video frames. Deep learning techniques like YOLO-V4, FlowNet, VGG-16, and CNN LSTM are used in the proposed architecture. The contributions of the research work are summarized as follows.

- A hybrid deep learning architecture for spatial feature processing is presented using YOLO-V4 and VGG 16.
- Temporal feature processing model is presented combining the optical FlowNet with VGG-16 architecture.
- Presented a feature fusion model using convolutional fusion to fuse spatial and temporal features obtained from hybrid YOLO-V4 and Optical FlowNet model.
- A hybrid deep learning architecture combining a convolutional neural network with Long Short Term Memory (LSTM) is presented to classify the optimal features of the fusion network.
- Presented an intense simulation analysis using a benchmark UCF crime dataset to evaluate the proposed model performances.
- Comparative analysis of the proposed model and recent anomaly detection models are presented to validate the better performance of the proposed model in terms of precision, recall, f1-score, and accuracy.

The remaining portion of the article is arranged in the following order. Section 2 summarizes the recent advancements in anomaly event detection based on existing literature works. The initial portion of section 3 presents the proposed detection model as an overview, followed by spatial feature processing using hybrid YOLO-V4 is presented. Later sections cover the temporal feature processing using hybrid optical FlowNet, feature fusion, and classification process. The experimental results are presented in section 4 followed by the conclusion in section 5.

2. RELATED WORKS

The challenges in abnormal event detection from the video are widely discussed by the research community. The non-deterministic definitions of abnormal events have insufficient training data, which makes the detection process more challenging. An adversarial learning procedure for abnormal event detection is reported in [4], which detects the events by differentiating the normal patterns based on the error function. An autoencoder is used to define the error function and it is combined with Generative Adversarial Networks (GAN) to improve the reconstruction ability. The attention model used in the presented work selects the informative parts for decoding and preserves the important features in the detection process. An optical flow-based two-stream video anomaly detection presented in [5] includes a convolutional autoencoder long-short term memory network to obtain raw data. For the

given raw data, the optical flow has been obtained to detect the anomalies considering the reconstruction error. Further, a weighted Euclidean loss is considered to concentrate more on network foregrounds. The presented approach performs global-local analysis to detect the anomalies in raw data.

Abnormal event detection using generative adversarial networks is presented in [6], which includes a super-resolution mechanism and a self-attention mechanism to form the network architecture. For the generator, an autoencoder is used, which includes a dense residual network and a self-attention mechanism. The discriminator model also includes self-attention based on the relativistic discriminator. Optical flow and gradient differences between frames are used to predict the final normal and abnormal events in the detection model. Video anomaly detection using dual discriminator-based generative adversarial networks has been presented in [7] to overcome the limitations in existing anomaly detection applications. The presented approach predicts the future frames using a generator, which is similar to ground truth. Using frame discriminator and motion discriminator, the generator model realizes the consecutive frames to confirm the originality of image frames. The originality of optical flow is determined using a motion discriminator so that real and fake optical flows can be sampled from original videos.

Recurrent Neural Network-based anomaly event detection is presented in [8], which considers the deep features of active learning, sparse representation, and dictionary learning. The motion fusion block used in the presented work eliminates the background noises and alleviates the data deficiency. Sparse representations are learned by the presented recurrent neural network, whereas dictionary learning is performed using an adaptive iterative hard thresholding algorithm. Video anomaly detection using generative adversarial networks is presented in [9], which includes shortcut inception modules and residual skips to improve the learning ability of the neural network. The training parameters are reduced in the presented approach by adopting asymmetric convolution instead of traditional convolution layers. Finally, a multi-scale U-Net model is employed to hold the essential features so that reconstruction error in the generator network is reduced, which increases the overall detection accuracy of the abnormal event detection process. Feature processing in anomaly event detection from surveillance video is presented in [10] and considers the spatial and temporal features to categorize normal and abnormal events. The U-Net architecture has been used as a detection model for spatial information processing, whereas convolutional LSTM is used to handle temporal information. Experimental results prove that the presented approach performs better than conventional U-Net-based detection models in terms of accuracy.

Abnormal event detection from video based on feature series extraction using a particle filter algorithm is presented in [11]. It extracts specific features to produce an alert. Then L2-norm extraction based on optical flow is used to represent the video features and then feature series are tracked using a particle filter. Based on the variations in the feature series and error tracking, abnormal events are detected in the presented approach. However, the accuracy of the presented approach is low compared to existing methodologies. In [12], a semi-supervised generative adversarial network-based video anomaly detection is presented, which learns the features in image space and latent space. In order to reduce the prediction errors, the frames for the future are predicted in the first step. Later, the predicted frames are encoded along with ground truths in latent space to minimize the differences. In the evaluation phase, the normal scores of each frame in the image and latent spaces are considered to capture the distributed information of the data, which improves the detection accuracy.

A multi-encoder single decoder network-based abnormal event model is presented in [13], which considers the encoding motions and cues individually and labels the event as normal and abnormal. The differences between raw frames and adjacent frames are obtained to define the content and motion sources. Improved learning ability and maximum running speed are the observed features of the presented detection model. Similar deep spatiotemporal translation network-based anomaly detection has been presented in [14], including edge wrapping and generative adversarial networks to detect anomalies. The presented approach performs a novel fusion of real optical flow frames with concatenated frames. Edge wrapping reduces the noise and suppresses the abnormal object's edges.

A stacked fully connected variational autoencoder and a skip convolutional variational

autoencoder has been presented in [15] to discriminate between normal and abnormal events in surveillance videos. The presented detection utilizes the key features of deep generative networks to attain better performance in local and global event detection. A similar stacked convolutional encoder is used in the anomaly detection model presented in [16] generates low-dimensional high-level features and trains the classifier for better performance. Meanwhile, to correct mapping relations and feature representations, a decoder unit is used, which reconstructs the samples from low-dimensional feature representations. Unlike traditional methods, the features are separated into two stages and automatically extracted using a neural network for better detection performance.

Abnormal event detection in the surveillance video is crucial as most of the anomaly events occur only for a short duration and the remaining part of the video has normal events. So, it is necessary to introduce a fast and simple detection method. Weakly-supervised anomaly detection based on deep learning networks is presented in [17]. It generates pseudo labels using binary clustering of temporal and spatial features. The presented approach enhances the detection accuracy through the network and clustering models compared to existing detection approaches. A similar weakly supervised anomaly detection model presented in [18] includes a cross-epoch learning procedure combined with a hard instance bank to enhance the training performance of the detection model. The high detection rate has been obtained using the cross-epoch learning approach. The presented approach can be integrated into existing detection frameworks for better performance improvement. From the literature analysis, it can be observed that most of the anomaly detection models are based on Generative Adversarial Networks (GAN), which include autoencoder and discriminator modules. However, the AUC performance of adversarial networks is not up to the mark and it can be improved if suitable feature processing techniques are included in the detection model. As of now, limited research models are presented as hybrid models, which include different architectures instead of GAN, but the performance of such hybrid models can also be improved if both spatial and temporal features are considered for analysis. Based on these observations, a novel hybrid deep learning architecture to handle spatial and temporal features from video keyframes to detect abnormal events is presented in the following section.

3. PROPOSED WORK

The proposed two-stream hybrid deep learning-based temporal and spatial features-based anomaly event detection from video keyframes is presented in this section. The dataset used in the proposed work is UCF crime data which has images extracted from the surveillance video. The data is prepared by extracting every 10th frame from a full-length video and combined for every video in that class. The images are in .png (Portable Network Graphics) format for all the classes. The process flow of the proposed model is depicted in figure 1. The process starts from the initial preprocessing of video frames which resize the images and selects the HSV features. HSV histogram is obtained based on the color features. HSV defines the elements Hue for H, Saturation in an image for S and Brightness in the image for V. In the proposed work, the hue elements are split into eight different parts, similarly saturation and brightness are split into three parts. Totally 72 subspaces (3x3x8) are obtained as HSV space. To derive the color components, different weights are assigned to HSV components which are based on human visual sensitivity. From this, a feature vector is represented in one-dimensional space. A similarity measure is obtained based on the ratio of object similar degree and overall objects. Considering the distance and similarity measure the keyframes are selected using k-means clustering. Followed by keyframe extraction, the selected keyframes are provided as input to the two-stream model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

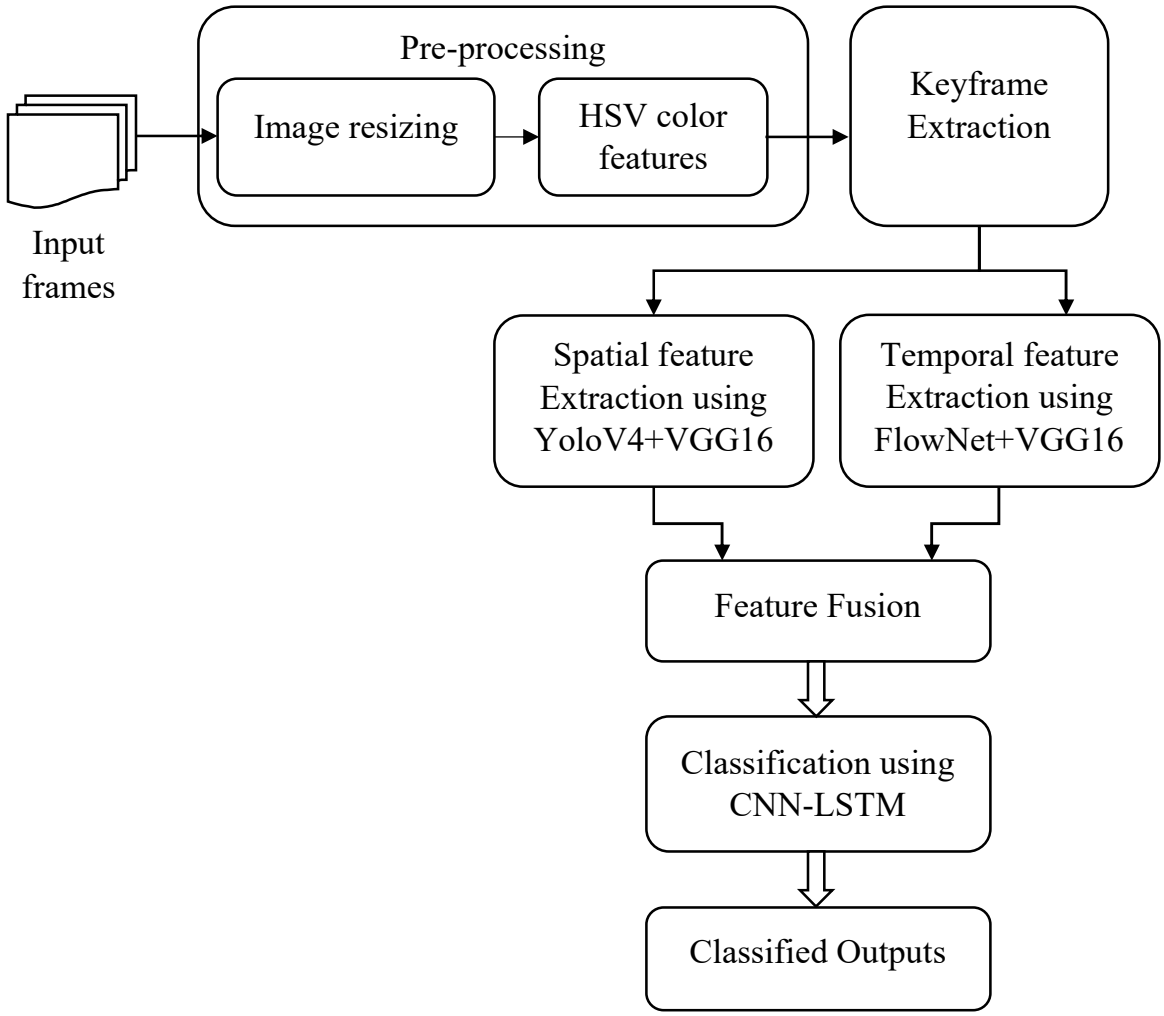


Figure 1 Overview Of The Proposed Model

In the two-stream model, temporal and spatial features are extracted separately using two different architectures. For spatial feature extraction, YOLOv4 is combined with VGG16 and for temporal feature extraction optical FlowNet and VGG16 are combined to select the optimal features. the selected features are concatenated or fused before classification. Finally, for the classification convolutional neural network -LSTM model has been used in the proposed architecture. The mathematical model for the proposed architecture is presented in this section. We have already performed the anomaly event detection and classification based on temporal features from keyframes [19], however, the accuracy is not up to the mark which motivated us to bring this proposed model. The major objective of this research work is to improve the detection rate and classification accuracy in anomaly event detection compared to existing methodologies. The novelty of the research work is present in the hybrid two-stream architecture which efficiently utilizes the deep characteristics and enhances the classification accuracy in the detection process.

3.1 Spatial feature processing using YOLOV4

Spatial feature extraction from video frames defines the relative spatial area of an object and its relationship with other objects. The major objective of spatial feature extraction is to differentiate the dynamic and static contexts in the keyframes. The static context defines the position of the moving object and the dynamic context defines the moving object along with the field. So,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

it is essential to understand the difference between static and dynamic contexts while processing spatial information. The proposed architecture includes YOLO-V4 for initial-level spatial feature extraction. You Only Look Once (YOLO) is a familiar object detection algorithm that generates object location coordinates and class probabilities. YOLO-V4 is an optimized model of YOLO-V3 in which the DarkNet53 is replaced with CSPDarkNet53 and it is considered the backbone of the network. The CSPDarkNet53 selects the optimum features so that the final classification accuracy has been improved while using YOLO-V4. The major reason for selecting YOLO-V4 over V3 is its processing ability of frames per second (FPS) which improves the average precision in real-time applications. Also, a better tradeoff has been obtained between detection speed and accuracy in YOLO-V4 compared to earlier versions.

The architecture of YOLO-V4 is composed of CSPDarknet53, Neck, and head in which the CSPDarknet53 is considered the network backbone. The neck includes Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN). The head portion is the final prediction part of the model. The CSPDarknet53 has a receptive field that includes multiple convolution layers which have numerous parameters. In order to increase the receptive field, spatial pyramid pooling is added in the YOLO-V4 which separates the significant context features without reducing the network operating speed. A path aggregation network is a form of parameter aggregation process which is used on different levels instead of a conventional feature pyramid network in YOLO-V3. The major objectives of using YOLO-V4 is to utilize its self-adversarial training features. this allows the detection process even if the objects are present outside the normal context. Also, the bath normalization provides different activation statistics for images on each layer which reduces the batch size in the processing. The self-adversarial network is a two-stage network in which the neural network in the first stage alerts the original image instead of network weights. While in the second stage the network is trained to detect the object.

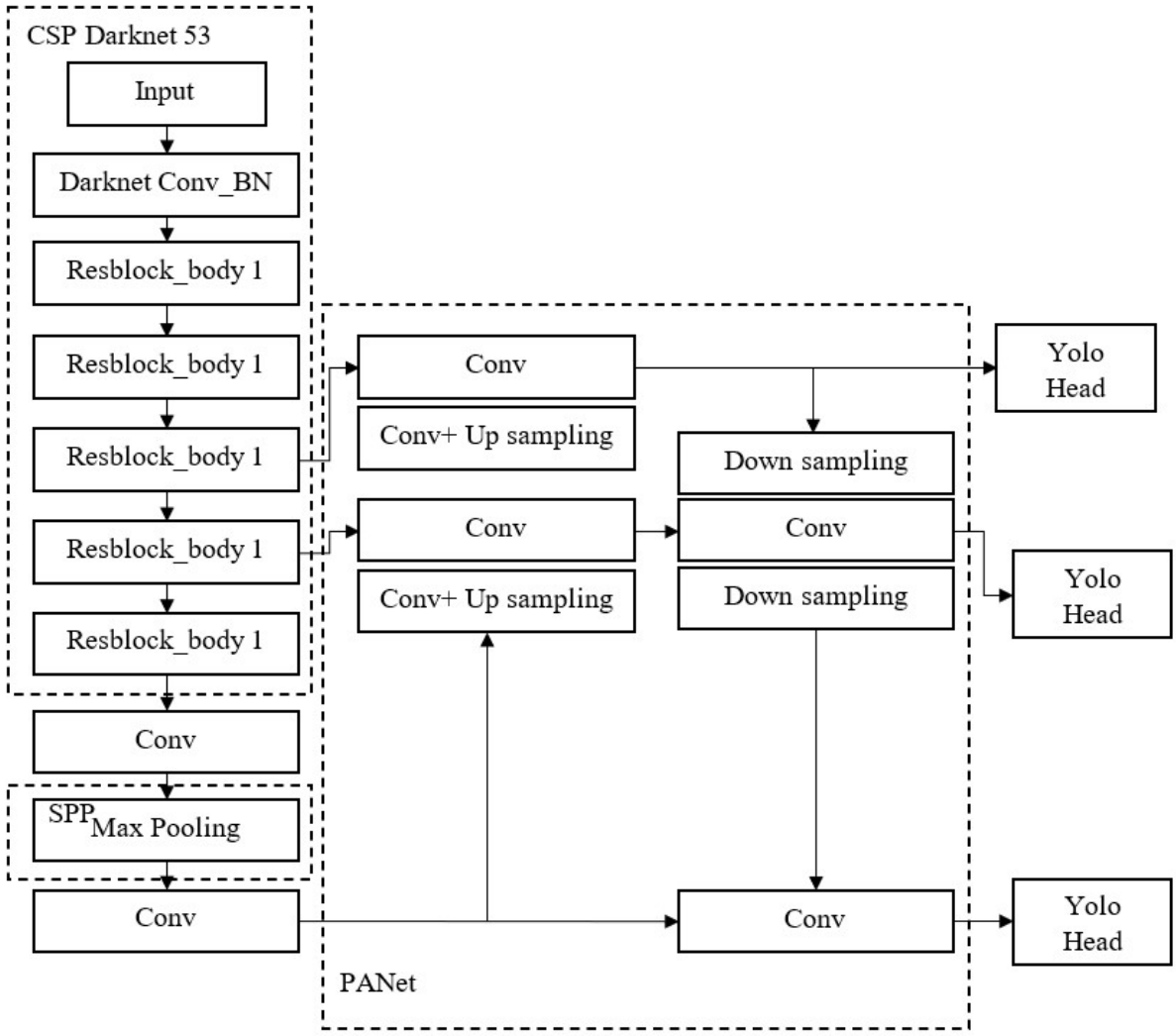


Figure 2 Architecture Of YOLO-V4 For Spatial Feature Processing

3.2 Temporal feature processing using Optical FlowNet

Most video semantics do not really happen in isolation, and a group of interest can be easily identified by how its semantics relates to its context. The earlier spatial features consider the object aspect ratio, principal axis direction, color or texture features, boundary descriptions, acceleration direction, etc., To simplify the definition, spatial features will capture the changes in the space due to the movement but in the case of temporal features, the time factors in the movement are considered. Temporal features help to recognize the object's movement in a video frame as an optical flow. The object motion pattern can be represented as an optical flow from stacked video frames and it can be used to define the motion structure, stabilization, and compression processes. However, modeling temporal optical flow for video processing is a critical task. So various metrics which are based on optical flow are used to extract the motion features from stacked video frames. Class descriptors such as Histograms of optical flow (HOF), dense trajectory, and histogram of oriented optical flow (HOOF) are explored by the research community to define the motion features. Similarly, conditional random fields, hidden Markov models, Bayesian networks, etc., are used to explore the long-term optical flow features. Recently, various deep learning models like a convolutional neural network, and Recurrent neural network are used to learn the optical flow features.

A Convolutional neural network-based optical flow method is considered for temporal feature processing in the proposed model. The proposed model is trained with motion displacement data which includes real-world action videos. Thus the feature extraction using optical FlowNet will effectively validate the human activities from a sequence of frames. Basically, FlowNet is designed for optical flow detection which performs the detection by training the network with consecutive frames. In the initial layers of the FlowNet model, the images have been separated using convolutional kernels to extract the image semantics. Followed by the convolution layer, the correlation layer is used to combine the image's convolutional features. This process is performed by multiplicative path comparisons between feature maps. The reason for selecting CNN-based optical flow is due to the feature benefits of CNN in image processing applications specifically the feature representation in motion patterns. Convolution layers in the CNN analyze the primitive features whereas the deeper layers with receptive fields will process the high-level semantics effectively.

The final convolutional layer of FlowNet is used to estimate the temporal features in which the action moves in the consecutive frames are deconvoluted to define the detected flow. The FlowNet model has been fed with two consecutive frames and the feature maps are extracted. For each feature map, global average pooling is applied to obtain the dense motion features. the extracted features represent the feature maps as global features. Global average pooling provides the entire feature map as a single value which includes the kernels. The major reason for utilizing global average pooling in FlowNet is to reduce the dimensions and to obtain the final temporal features, the results of global average pooling are combined layer-wise which represents the final temporal features. Figure 3 depicts the CNN based FlowNet architecture used in the proposed model for temporal feature extraction.

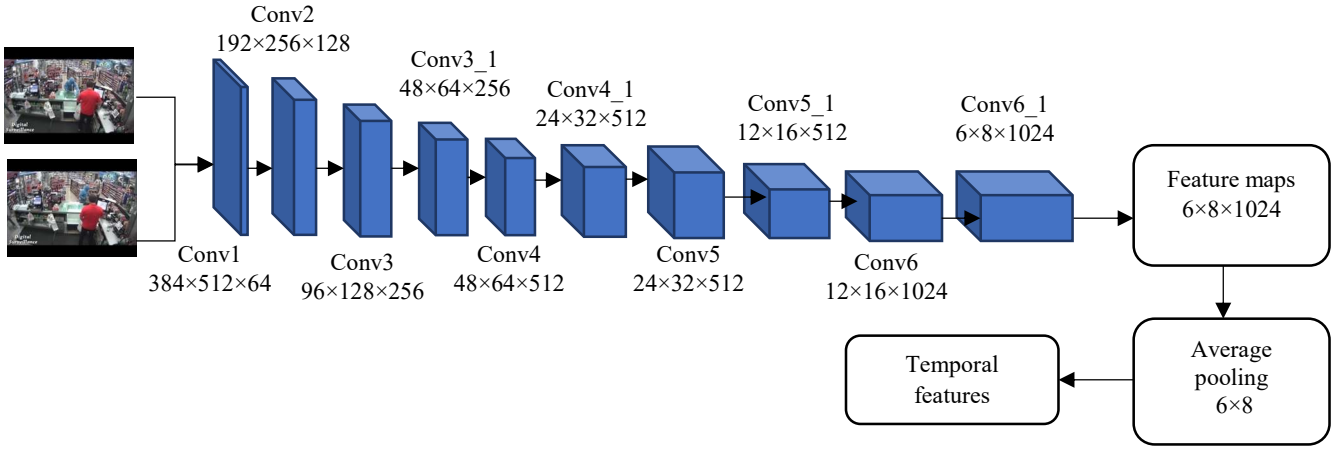


Figure 3 CNN Based Flownet Model

The CNN initial layers analyze small primitive features whereas deep layers extract the high-level motion features. In the proposed feature extraction process the final layer of the FlowNet model is included for flow estimation. The consecutive frames and their motions are represented in this layer to detect the flow. As shown in the architecture given in figure 3, two consecutive frames are provided as input to the CNN-based optical flow Net model and the feature maps are extracted from the final convolution layer. For each feature map, an average pooling is applied to obtain the motion feature which is considered a global feature map. In order to reduce the feature map dimensions, global average pooling is employed, and to obtain final temporal feature vectors, all the average pooling results are layer wisely combined.

3.3 Optimal feature selection using VGG-16 (Temporal and Spatial)

VGG-16 is a famous CNN model that is being used for ef□cient feature computation. The size of the input image used for VGG-16 is 224 × 224 × 3. This network can be used for RGB images also. The architecture of VGG-16 is based on the input layer, 5 layers of the max pool,

ve segments of convolution layers having 13 layers of the total, and 3 Fully Connected (FC) layers. The lters of 3 × 3 are used in the rst two convolutional layers. The lter of size 3 × 3 with stride 1 is used. A total of 64 lters are used in the rst two layers and it gives an output of 224 × 224 × 64. After that, the pooling layer is used, and it gives an output of 112 × 112 × 64. After the pooling layer, two more convolution layers are used having a total 128 of lters and it gives us the output size of 112 × 112 × 128. After these convolution layers pooling layer is used, and it gives the output of size 56 × 56 × 128. After this pooling layer, two more convolution layers are added having a 256 lter. Then pooling layer is added. After that, 3 convolution layers are added having lters of 512. After this convolution layer, the pooling layer is added again. After that, 3 convolution layers are added having lters of 512. Then pooling layer is added again after these convolution layers. Finlay, the fully connected layers are added, and the nal output size is 7 × 7 × 512 into the FC layer. There is a total of three FC layers. The total channels at the rst two FC layers are 4096 and at the third FC layer is 1000. ReLU activation function is used in all hidden layers. The architecture of the VGG-16 is illustrated i

The proposed anomaly detection architecture for spatial and temporal feature processing includes VGG-16 for optical feature selection in the second stage. In the first stage, network learns the spatial features using YOLO-V4 and temporal features using the optical FlowNet model. To enhance the feature processing ability and enhance accuracy, optimal features are selected from the existing extracted features. VGG16 is familiar for feature computation which has the ability to process RGB images. The architecture of VGG16 has an input layer, max pool, and five segments of convolution layers which have 13 layers in total. Followed by convolution layers three fully connected layers are included in the architecture. The initial two convolutional layers have 3×3 filters with stride rate 1 and a total of 64 filters are used in the initial layers. Followed by the initial two convolutional layers, the pooling layer is used with 128 filters. After the pooling layer, two more convolutional layers are added with 256 filters. The pooling layer is included after this convolutional layer. Again, three convolution layers have been added with 512 filters and a pooling layer is added after that. Similarly, again three convolutional layers have been added with the pooling layer two times and finally, fully connected layers are added. ReLU activation function is used and the complete architecture of VGG16 is presented in figure 4.

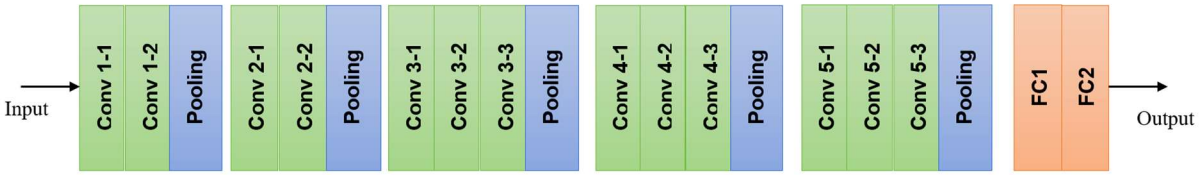


Figure 4 VGG-16 architecture

In the optimal deep feature extraction using VGG-16, the features are computed in the fully connected layers. The initial feature extraction in VGG-16 is carried out by transfer learning so that the network model is trained with motion data on various angles and the activation of fully connected layers are performed in parallel to handle the spatial and temporal features.

3.4 Feature Fusion

Generally, the spatial and temporal feature fusion has been performed in the recognizable layers i.e., Fully connected layers. However, pixel-level feature fusion cannot be performed by conventional learning and training processes. If the action involves a specific object in motion over periodically without fusing the motion information and spatial information the network fails to detect the targeted information. Due to this reason, feature maps on each channel must be fused at the same position. However, this kind of fusion is possible only if the motion information and spatial information have a similar network structure. In the case of multiple channels that handle the temporal and spatial features, it is essential to determine the relationship between them. It can be assumed that spatial networks are responsible for extracting features from different regions at the same time temporal network extracts the motion features from different regions. This will make the fusion process more complex, to overcome this, the relationship between channels must be determined. Considering this limitation in the traditional feature fusion process, a spatial-

temporal feature fusion is presented using the convolution process. The feature fusion process is mathematically expressed as follows.

$$y_t = f(x_t^a, x_t^b) \quad (1)$$

where the fusion function is represented as f, t represents the time and the spatial network feature map is represented as x_t^a. The temporal feature map is represented as x_t^b and y_t represents the output feature map after fusion process. The feature maps are represented based on the width w, height h and number of channels D. In the first step the features are cascaded and stacked on the fusion layer which is mathematically represented as

$$y_t^c = f(x_t^a, x_t^b) \quad (2)$$

Considering the spatial position i, j of the feature map the stacking process is formulated as follows.

$$y_{i,j,2d}^c = x_{i,j,d}^a \cdot y_{i,j,2d-1}^c = x_{i,j,d}^b \quad (3)$$

In order to make the network to obtain the relationship between temporal and spatial channels through learning, convolutional fusion is carried out which is formulated as follows.

$$y^{conv} = y_t^c * f + b \quad (4)$$

where the filters are represented as f and its dimensions are given as $n \times m \times 2D$. Using the filters, the dimensionality of the channels is reduced so that accuracy in the anomaly detection is increased. Figure 6 depicts the LSTM network model in detail which is used in the feature classification architecture given in figure 5.

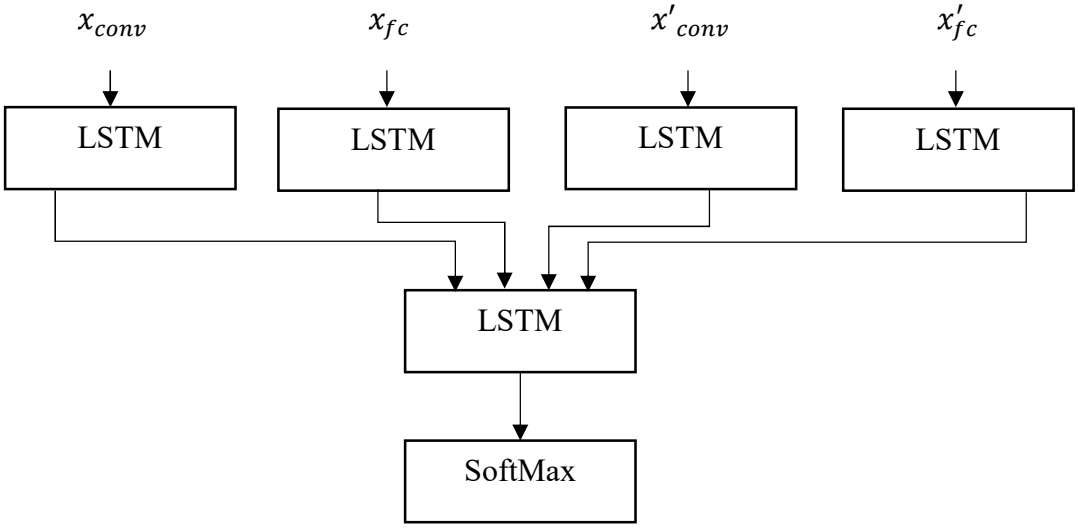


Figure 5 Feature classification using CNN-LSTM

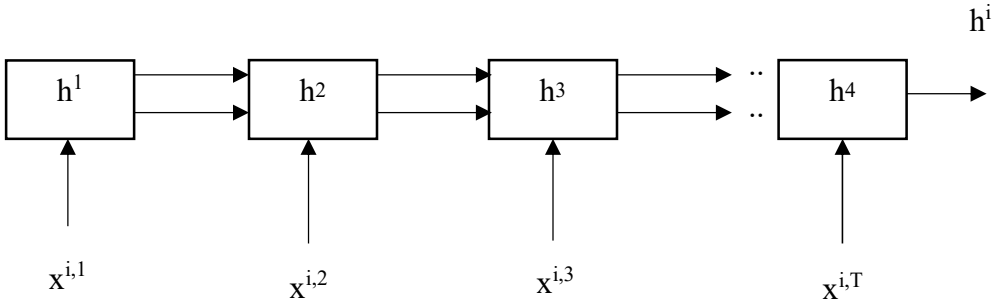


Figure 6 LSTM Networks Used In The Proposed Network

3.5 Feature classification using CNN-LSTM

For feature classification i.e., anomaly detection, a combination of convolutional neural network (CNN) and Long Short-Term Memory (LSTM) algorithms are used in the proposed model. Figure 5 depicts the proposed classification model in which the feature map after fusion is represented as x_{conv}, whereas the first fully connected layer is represented as x_{fc}. The feature map after weighted pooling is represented as x'_{conv} and the fully connected network after weighted pooling is represented as x'_{fc}. The LSTM network is a cycle network which handles the temporal features in input and output.

The proposed multilayer LSTM has an input unit and output unit which handles the sequences of input features. The hidden states of LSTM handle the input sequence and realize the video sequence representations to acquire the essential information. The end-to-end back propagation realize the information exchange between the convolutional layers $(x_{conv}/x'_{conv} - LSTM)$ and fully connected layers $(x_{fc}/x'_{fc} - LSTM)$. Once the information exchange is finished the back regional attention is performed meanwhile the backpropagation training is improved. The mathematical model for LSTM is given as follows. In the first layer, features from convolutional layers and fully connected layers are fused using LSTM to realize the video frames which is formulated as follows.

$$h_{conv}^{i,t} = LSTM\left(x_{conv}^{i,t}, h_{conv}^{i,t-1}\right) \quad (5)$$

$$h_{conv}^{i} = \left[h_{conv}^{i,1}, h_{conv}^{i,2}, \dots, h_{conv}^{i,T}\right] \quad (6)$$

$$h_{fc}^{i,t} = LSTM\left(x_{fc}^{i,t}, h_{fc}^{i,t-1}\right) \quad (7)$$

$$h_{fc}^{i} = \left[h_{fc}^{i,1}, h_{fc}^{i,2}, \dots, h_{fc}^{i,T}\right] \quad (8)$$

The first layer outputs are again provided to LSTM and SoftMax classifier to obtain the results. Mathematically the process is expressed as follows.

$$h^i = LSTM\left(W\left[h_{conv}^{i,t}, h_{fc}^{i,t}\right]\right) \quad (9)$$

$$y^i = SoftMax(h^i) \quad (10)$$

Thus, the final classified status defines the anomaly status from the video key frames. The video frames sampling time is considered as $t, t+\tau, \dots, T\tau$.

4. RESULTS AND DISCUSSION

The proposed classification model for anomaly event detection using hybrid deep learning techniques is verified through simulation analysis performed using Python, which is installed on the Linux platform version 18.04 Ubuntu. The processing unit comprises an Intel i5 processor with 8GB of RAM. The Benchmark UCF crime dataset [20] is used in the simulation analysis and the details of the dataset are provided in the table 1. The dataset contains 1900 real world surveillance videos under 13 different categories like Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. The proposed model experimentation considers only

seven categories like abuse, arson, fighting, robbery, shooting, stealing, and vandalism. The total number of frames in the UCF crime dataset for each category is illustrated in figure 7.

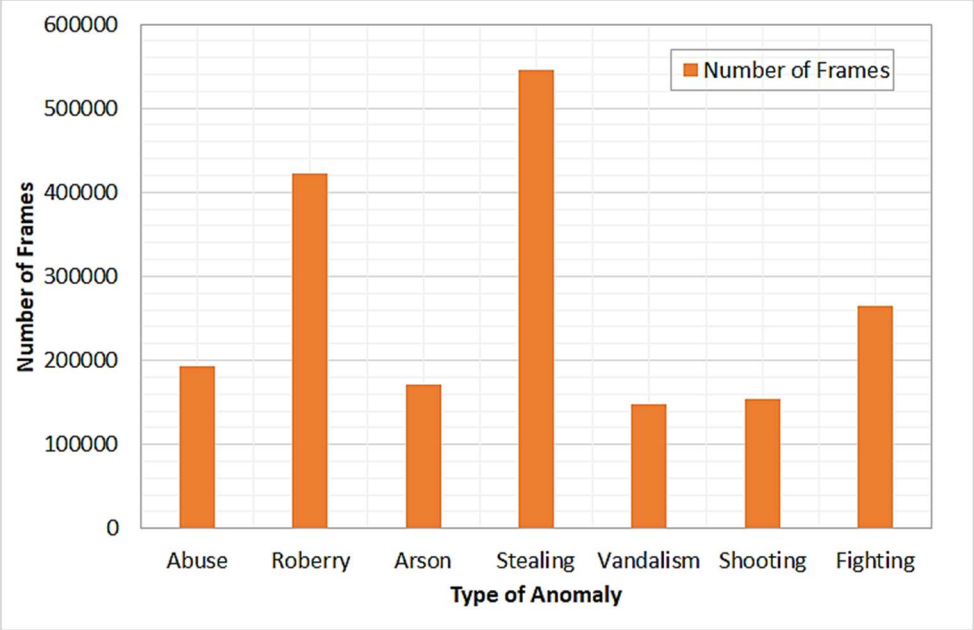


Figure 7 Frames in UCF crime dataset

Table 1 Training and testing data from UCF crime Dataset

S.No	Category	Actual frames	Train	Test
1	Abuse	193497	154798	38699
2	Robbery	422594	338075	84519
3	Arson	171861	137489	34372
4	Stealing	545757	436606	109151
5	Vandalism	147142	117714	29428
6	Shooting	154692	123754	30938
7	Fighting	264547	211638	52909

The keyframe extracted for each category is depicted in figure 8 as a comparative analysis with actual frames. It can be observed that the extracted frames are the necessary frames which are used to detect the anomalies. This process reduces the overall computation time of the classifier and increases its accuracy.

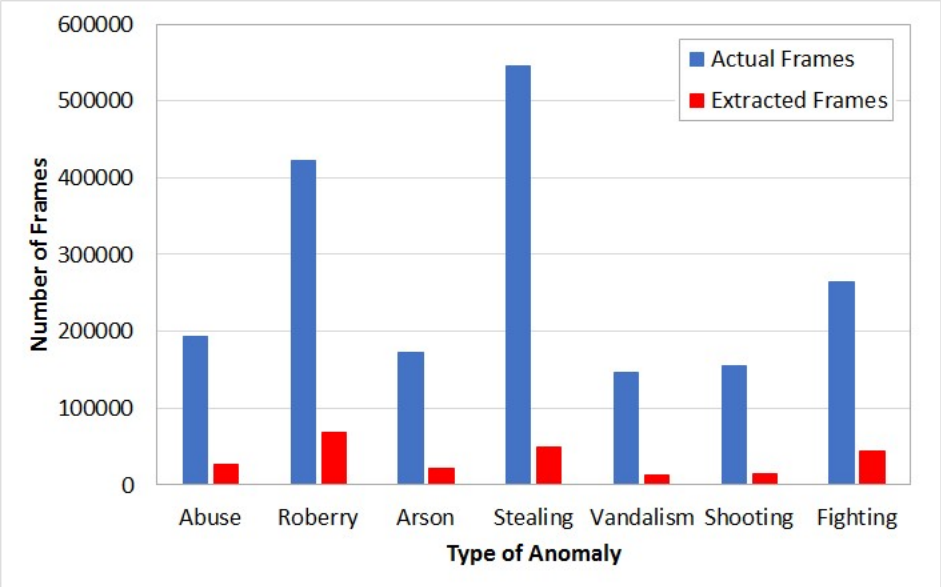


Figure 8 Comparison Of The Actual Frame And Extracted Frames In The Proposed Approach

Figure 9 depicts the sample keyframes for Robbery, Arson, Stealing, Vandalism, Shooting and Fighting.

(a)

(b)

(c)

(d)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(e)

(f)

Figure 9 Keyframes Extracted From UCF Crime Dataset (Top To Bottom) (A) Robbery (B) Arson (C) Stealing (D) Vandalism (E) Shooting(F) Fighting

The final classification results of the proposed anomaly detection are presented in Figure 10, which detects the people in the key frames. This detection is performed based on the optimal features obtained from hybrid deep learning models used for spatial and temporal feature extraction. Fusing the temporal and spatial features using a convolution process improves the detection accuracy.

(a)

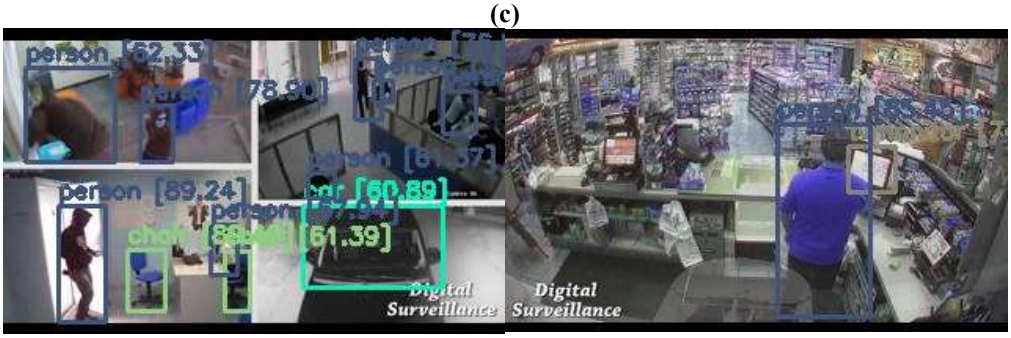
(b)

Journal of Theoretical and Applied Information Technology

30th September 2023. Vol.101. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(c)



(d)

(e)

(f)

Figure 10 Final results of proposed anomaly detection model (a) Arson (b)Fighting (c) Stealing (d) Robbery (e) Shooting (f) Vandalism

Figure 11 depicts the receiver operating characteristics (ROC) of the proposed model based on the true positive and false positive rates. It can be observed that the false positive rate is at its minimum and it gradually increases. Even with a high false positive rate, the variation in true

positive rate is minimal. This indicates the better performance of the proposed anomaly detection model.

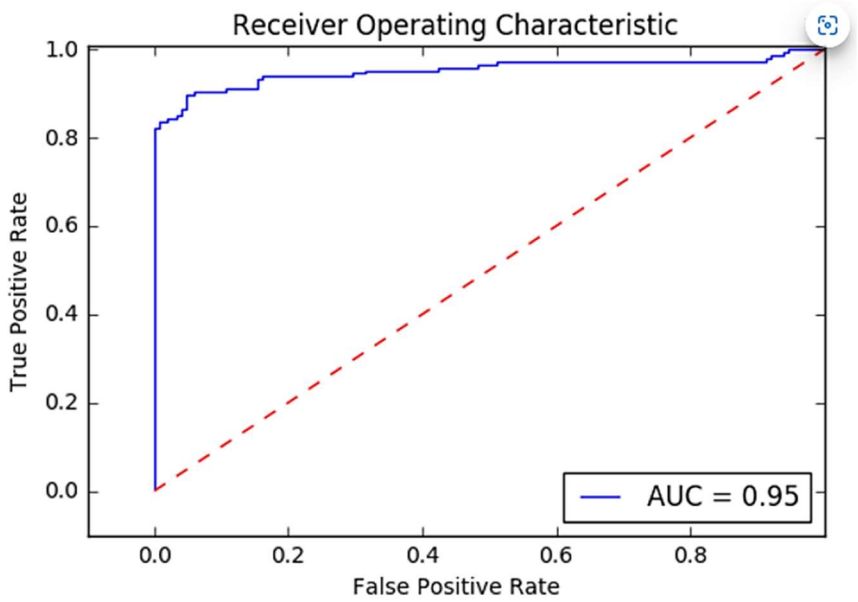


Figure 11 Receiver Operating Characteristics (ROC) Analysis

The training and testing of the proposed model is performed in the ratio of 80:20, and the number of samples used for training and testing is depicted in table 1. Figure 12 depicts the details of the accuracy and loss of the proposed model for 100 epochs. After 100 epochs, the accuracy and loss remain the same and don't show any improvements. Based on the results, the accuracy of the proposed model has been fixed at 100 epochs as 95.6%.

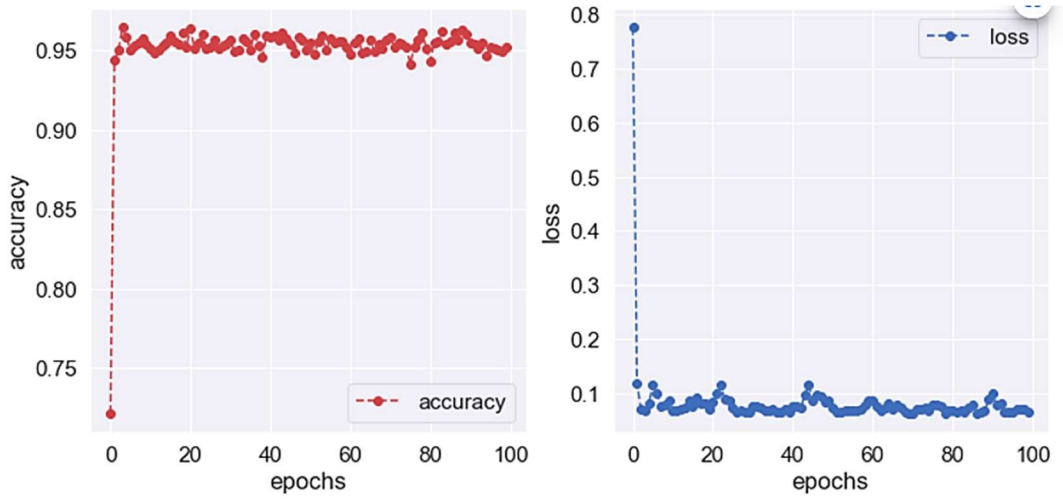


Figure 12 Accuracy and Loss analysis

Based on the true positive, false positive, true negative, and false negative values obtained in the testing process, the performance of the proposed model is further evaluated using performance metrics like recall, precision, and f1-score. Also, to validate the superior performance of the proposed model, existing methods for anomaly detection like Mobile Net V2 + LSTM and Mobile Net V2 + BD-LSTM [20],CLSTM, CNN-RNN[21] are considered for comparative analysis.

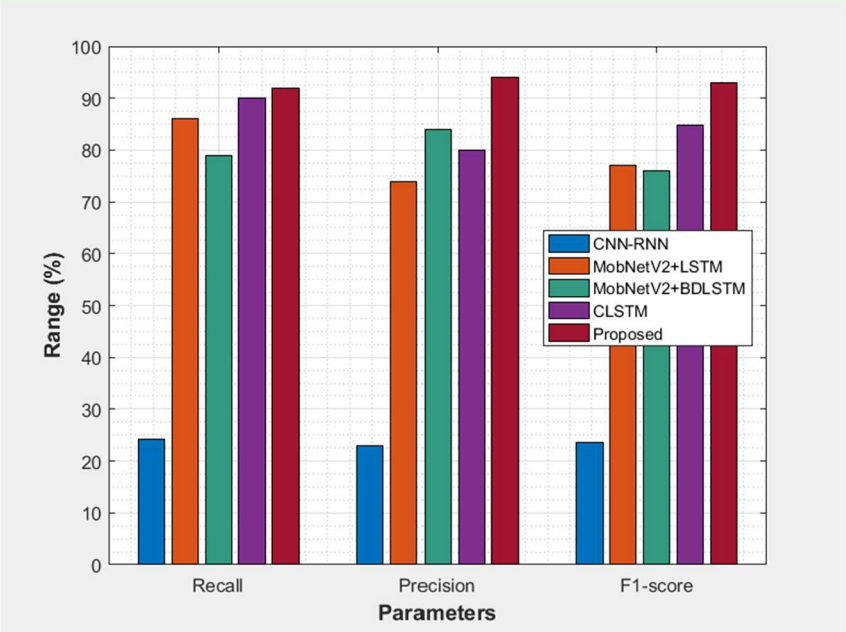


Figure 13Performance Comparative Analysis

Figure 13 depicts the comparative analysis of the proposed model and the existing model for precision, recall, and f1-score metrics. It can be observed from the results that the proposed model attains maximum values for the metrics due to the optimal feature selection and processing. The spatial and temporal feature processing reduces false values while improving performance in the anomaly detection process. Due to the improper feature handling the performance of hybrid CNN-RNN model exhibits poor results compared to remaining methodologies.

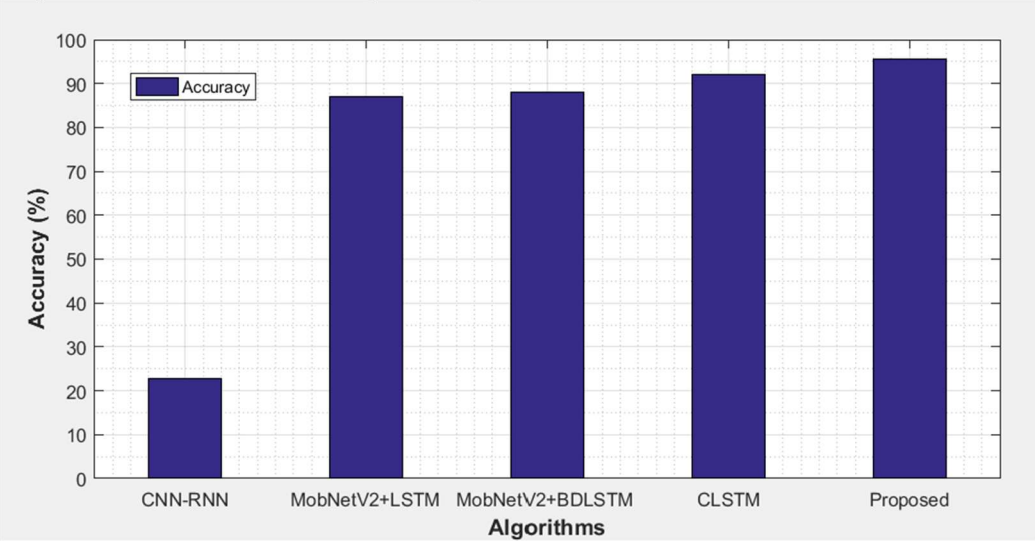


Figure 14Accuracy analysis

The accuracy analysis of the proposed model and existing models is depicted in figure 14. The results clearly show that the maximum performance is attained by the proposed model. The obtained 95.6% accuracy of the proposed model is approximately 4% greater than the CLSTM based model. The Mobile Net V2 +LSTM and Mobile Net V2 +BD-LSTM obtain 87% and 88% accuracy, which is approximately 9% less than the proposed model. The reason for the proposed model's maximum accuracy is the efficient processing of temporal and spatial features using optical FlowNet and YOLO-V4. The convolution fusion enhances the accuracy by

combing the spatial and temporal features, and finally, the hybrid classification model effectively classifies the features and detects the anomalies.

Table 2 Comparative analysis

S.No	Methods	Results
1	Convolutional autoencoder[23]	50.6%
2	Sparse combination learning[24]	65.51%
3	Deep MIL [20]	75.41%
4	VGG-16 [25]	87.27%
5	Inception V3 - VGG-16 [26]	88.74 %
6	Inception V3 [25]	94.54%
7	Proposed Method	95.6%

To present a comprehensive comparative analysis of proposed model, existing research works which utilizes the same dataset are considered for analysis and its common parameters like AUC and accuracy are used to compare with proposed model results. The maximum results obtained by the proposed model indicates the better performances in abnormal event detection.

5. CONCLUSION

This research work presents hybrid deep learning approaches for video anomaly detection and classification using YOLO-V4, FlowNet, VGG-16, and CNN-LSTM. Instead of considering only the spatial or temporal features in traditional anomaly detection approaches, the proposed model considers both the spatial and temporal features to attain maximum accuracy. Three hybrid architectures are used in the proposed model for feature processing and classification. The spatial features from the keyframes are processed using YOLO-V4 and VGG16. The temporal features are obtained using FlowNet with VGG16. The obtained features are fused using a convolutional fusion process and finally classified through a hybrid deep learning model that includes CNN and LSTM approaches. The final classified results provide information about the anomalies in the keyframes. The minor limitation of the research work is its multiple hybrid deep learning architecture. However, considering the best performance, this can be neglected. In the future, this research work can be extended by using multiple deep learning networks as a comparative analysis.

REFERENCES

[1] Arunnehru.J, and M. Kalaiselvi Geetha (2016), "Vision-Based Human Action Recognition in Surveillance Videos Using Motion Projection Profile Features.", Lecture Notes in Computer Science (LNCS), vol.9468, no.1, pp. 460-471. 2016.

[2] Wenqing Chu, Hongyang Xue, Chengwei Yao, Deng Cai (2019), "Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos", IEEE Transactions on Multimedia, vol. 21, no. 1, pp. 246-255.

[3] Rajaram, Santhoshkumar, Kalaiselvi Geetha.M(2019), "Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks", Procedia Computer Science, vol.152, pp.158-165.

[4] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, Yunde Jia (2020), "Learning Normal Patterns via Adversarial Attention-Based Autoencoder for Abnormal Event Detection in Videos", IEEE Transactions on Multimedia, vol. 22, no. 8, pp. 2138-2148.

[5] Biao Yang, Jinmeng Cao, Nan Wang, Xiaofeng Liu (2019), "Anomalous Behaviors Detection in Moving Crowds Based on a Weighted Convolutional Autoencoder-Long Short-Term Memory Network", IEEE Transactions on Cognitive and Developmental Systems, vol. 11, no. 4, pp. 473-482.

[6] Weichao Zhang, Guanjun Wang, Mengxing Huang, Hongyu Wang, Shaoping Wen (2021), "Generative Adversarial Networks for Abnormal Event Detection in Videos Based on Self-Attention Mechanism", IEEE Access, vol. 9, pp. 124847-124860.

[7] Fei Dong, Yu Zhang, Xiushan Nie (2020), "Dual Discriminator Generative Adversarial Network for Video Anomaly Detection", IEEE Access, vol. 8, pp. 88170-88176.

[8] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, Rick Siow Mong Goh (2019), "AnomalyNet: An Anomaly Detection Network for Video Surveillance", IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2537-2550.

[9] Savath Saypadith, Takao Onoye (2021), "An Approach to Detect Anomaly in Video Using Deep Generative Network", IEEE Access, vol. 9, pp. 150903-150910.

[10] Yuanyuan Li, Yiheng Cai, Jiaqi Liu, Shinan Lang, Xinfeng Zhang (2019), "Spatio-Temporal Unity Networking for Video Anomaly Detection,", IEEE Access, vol. 7, pp. 172425-172432.

[11] Xinwen Gao, Guoyao Xu, Shuaiqing Li, Yufan Wu, Edvins Dancigs, Juan Du (2019), "Particle Filter-Based Prediction for Anomaly Detection in Automatic Surveillance", IEEE Access, vol. 7, pp. 107550-107559.
[12] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, Yilong Yin (2021), "Normality Learning in Multispace for Video Anomaly Detection", IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 9, pp. 3694-3706.
[13] Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, Feng Yang (2021), "Multi-Encoder Towards Effective Anomaly Detection in Videos", IEEE Transactions on Multimedia, vol. 23, pp. 4106-4116.
[14] Thittaporn Ganokratanaa, Supavadee Aramvith, Nicu Sebe (2020), "Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network", IEEE Access, vol. 8, pp. 50312-50329.
[15] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, Chang Choi (2019), "Generative Neural Networks for Anomaly Detection in Crowded Scenes", IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1390-1399.
[16] Peng Wu, Jing Liu, Fang Shen (2020), "A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes", IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 7, pp. 2609-2622.
[17] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, Seung-Ik Lee (2020), "A Self-Reasoning Framework for Anomaly Detection Using Video-Level Labels", IEEE Signal Processing Letters, vol. 27, pp. 1705-1709.
[18] Shenghao Yu, Chong Wang, Qiaomei Mao, Yuqi Li, Jiafei Wu (2021), "Cross-Epoch Learning for Weakly Supervised Anomaly Detection in Surveillance Videos", IEEE Signal Processing Letters, vol. 28, pp. 2137-2141.
[19] Mangai. P, Geetha. M, K. Kumaravelan. G (2022), "Temporal Features-Based Anomaly Detection from Surveillance Videos using Deep Learning Techniques", 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 490-497.
[20] Waqas Sultani, Chen Chen, Mubarak Shah (2018), "Real-world Anomaly Detection in Surveillance Videos" Cornell University Library, arXiv:1801.04264 [cs.CV], vol.1. pp.1-10.
[21] Waseem Ullah, Amin Ullah , Tanveer Hussain, Zulfiqar Ahmad Khan and Sung Wook Baik (2021), "An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos", Sensor, pp. -17.
[22] Soheil Vosta and Kin-Choong Yow (2022), "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras", Applied sciences, vol.12, pp.1-15.
[23] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S (2016), "Learning temporal regularity in video sequences", In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 733–742.
[24] Lu, C., Shi, J., Jia, J. (2013), " Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision". pp. 2720–2727.
[25] Koppikar, U., Sujatha, C., Patil, P., & Mudenagudi, U (2020), "Real-World Anomaly Detection Using Deep Learning", Advances in Intelligent Systems and Computing, pp.333–342.
[26] Majhi, S., Dash, R., Sa, P.K. (2020), "Two-Stream CNN Architecture for Anomalous Event Detection in Real World Scenarios", In: Nain, N., Vipparthi, S., Raman, B. (eds) Computer Vision and Image Processing. CVIP 2019. Communications in Computer and Information Science, vol 1148, pp. 343–353.