Video Anomaly Classification Using Convolutional Neural Network with Bidirectional Long Short-Term Memory Using Spatio-temporal Adaptive Transformer

Divya Uluvaru Hoovayya [1,*], Josephine Prem Kumar [2], and Heena Kousar [1]

[1] Department of Computer Science and Engineering, East Point College of Engineering and Technology, Visvesvaraya Technological University, Belagavi, Karnataka, India
[2] Department of Computer Science and Engineering, Cambridge Institute of Technology, Visvesvaraya Technological University, Belagavi, Karnataka, India
Email: divyauhgopal@gmail.com (U.H.D.); d_prem_k@yahoo.com (J.P.M.); hkheenakousar73@gmail.com (H.K.)
*Corresponding author

***Abstract*—In smart cities, surveillance systems are extensively deployed to monitor public areas and critical infrastructures. These continuous video streams are a valuable source for analyzing real-time activities and supporting tasks such as object detection, behavior analysis, and incident monitoring. Among these applications, detecting unusual or suspicious events is particularly crucial for ensuring safety and enabling timely responses to threats. Because of its importance in enhancing urban security, video anomaly detection has emerged as a central research focus in the broader field of intelligent video analysis. Thus, this research proposes a new integrated framework of Convolutional Neural Networks with Bidirectional Long Short-Term Memory using Spatio-Temporal Adaptive Transformer (CNN-BiLSTM-STAT) named for the classification of video anomalies in surveillance cameras. The proposed CNN-BiLSTM-STAT was implemented by using different datasets, such as the University of Central Florida Crime (UCF-Crime) dataset, Real-Life Violence Situations (RLVS) dataset, Extreme Dataset for Violence Detection (XD_Violence), and Real-World Fight 2000 Dataset (RFW-2000). Then, image resizing and label encoding techniques are used in the preprocessing phase to improve the input data. Finally, the proposed integrated classifier CNN-BiLSTM-STAT was used to classify video anomalies into multiple classes. The experimental results demonstrate that the proposed CNN-BiLSTM-STAT method attains an optimal accuracy of 99.91% on the UCF-Crime dataset compared to the existing methods, such as Recurrent Neural Networks with LSTM (RNN-LSTM) and MobileNet.**

***Keywords*—bidirectional long short-term memory, convolutional neural network, spatio-temporal adaptive transformer, surveillance cameras, video anomalies**

I. Introduction

Video surveillance has become a critical component in the planning, operation and long-term sustainability of both urban and industrial environments [1]. It enhances the safety and security of infrastructure, public spaces and operational activities by providing continuous monitoring [2]. With rapid growth in urban development and industrial expansion, the use of Closed-Circuit Television (CCTV) systems to cover both large-scale and high-rise assets has increased dramatically [3]. However, expecting human operators to monitor and accurately analyze every video feed around the clock is both unrealistic and impractical [4]. Unusual events such as fights, road accidents, theft, abuse and violent crimes are rare and unpredictable, which makes them difficult to detect in real time through manual observation [5]. This highlights the urgent need for intelligent surveillance systems capable of automatically detecting and identifying such anomalies using advanced video-processing techniques [6, 7]. In critical situations, intelligent surveillance systems can provide quick alerts to authorities to take necessary action. Manual monitoring systems are often limited by human fatigue, cognitive overload and limited attention span [8]. In recent years, several computer vision-based approaches have been explored to detect, track and recognize objects such as people, vehicles and animals in surveillance footage, with the broader objective of understanding behaviors and identifying abnormal patterns efficiently [9, 10].

Detecting anomalies in complex environments, particularly crowded or high-traffic areas, is a major focus of modern video surveillance systems [11]. Anomalies refer to irregular or unexpected events that differ significantly from typical behavioral patterns [12]. Identifying and modelling such rare occurrences can be extremely challenging owing to their unpredictable nature and limited visibility in large-scale scenarios [13].

Manuscript received October 6, 2025; revised December 8, 2025; accepted January 26, 2026; published April 28, 2026.

Therefore, designing a reliable anomaly detection system is crucial for achieving intelligent and automated video surveillance [14]. Traditionally, anomaly detection in Industrial Control Systems (ICS) has relied on predefined rules and threshold values defined by the domain experts [15]. Although effective to some extent for known issues, these methods struggle to handle unfamiliar or evolving threats. Furthermore, manual monitoring and rule-based detection are labor-intensive and often result in high false alarms or missed detections, which makes them unsuitable for fast-paced and complex industrial setups [16, 17]. With recent advancements in Artificial Intelligence (AI), particularly in computer vision and Deep Learning (DL), significant improvements have been made in video analysis and behavior recognition [18]. Studies have shown that combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) significantly enhances the effectiveness of supervised models, particularly in detecting violent or abnormal activities within the video streams [19, 20].

In modern surveillance systems, identifying anomalous activities from continuous video streams remain as a significant challenge because of the dynamic nature of real-world environments, varied illumination, crowded scenes and unpredictable behavioral patterns. An important aim of this study is to propose a CNN with Bidirectional Long Short-Term Memory using Spatio-Temporal Adaptive Transformer named (CNN-BiLSTM-STAT) approach, which effectively classifies the video anomalies by integrating spatial and temporal attention mechanisms.

The key notes of this manuscript are as follows:

- The CNN-BiLSTM-STAT approach is proposed for classifying the different classes of video anomalies, which supports the enhancement of model performance.
- This research utilized different preprocessing techniques, such as image resizing and label encoding, to enhance image excellence.
- The significance of the proposed approach was validated using various performance measures based on different video anomaly datasets.

The remainder of this portion is arranged as trails. Section II describes the literature survey. The proposed methodology is described in Section III. Section IV presents the results and discussion, and Section V concludes the paper.

II. Literature Survey

In this section, the existing works related to video surveillance are discussed based on the DL approaches, along with their advantages and disadvantages.

A. Reconstruction-Based and Statistical Approaches

Ali [21] implemented a surveillance framework that combined Background Subtraction (BS), convolutional autoencoders, and object detection. This framework modeled each pixel using a Mixture of Gaussians (MoG) to isolate moving foreground objects. The extracted Regions of Interests (ROIs) were processed by autoencoders to distinguish abnormal events from normal activities in real time. However, the BS technique generated numerous false positives, as even normal scene changes were misclassified, thereby affecting the detection reliability.

Liu *et al.* [22] developed a Stochastic Video Normality (SVN) approach that learned local appearance patterns through deterministic multi-task learning and modeled global motion in a stochastic manner. They used recurrent networks to estimate motion via a conditional Gaussian distribution and introduced a masked autoencoder to reinforce the learning of spatio-temporal patterns. However, this approach failed to effectively handle the unique distinction between static and dynamic content, thereby making motion inference more complex.

B. Hybrid Models for Temporal Modeling

Jebur *et al.* [23] implemented an ensemble model in which CNN and LSTM were utilized for feature fusion to integrate diverse DL models for improved feature representation. The Gradient-weighted Class Activation Mapping (Grad-CAM) approach was employed as an interpretability technique to visualize crucial regions within images. The Grad-CAM leveraged gradients to effectively highlight the areas of the image that exerted significant influence over the decision-making process. Multi-task classification was utilized to enable the generalization of classifiers across various tasks. However, the Grad-CAM technique struggled with precise localization in complex or noisy images, which potentially led to less accurate visualization of important features.

Qasim and Verdu [24] developed an anomaly detection system using a deep CNN and a Simple Recurrent Unit (SRU). They utilized ResNet to extract spatial features from video frames, while the SRU captured temporal patterns because of its high parallelizability and expressive recurrence. The system succeeded in identifying and classifying unusual actions. Nevertheless, the application of early stopping during training reduced the overall performance and learning capacity of the model.

Kotkar and Sucharita [25] proposed a Modified Spatio-Temporal (MST) technique that extracted interest points from video frames. Their method began with Gaussian filtering for normalization, followed by motion tracking and cuboid generation. Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA) were used for dimensionality reduction before the extracted features were fed into an RNN with LSTM (RNN-LSTM) classifier. Nonetheless, redundant or noisy frames increased the computational demands and compromised model efficiency.

C. Lightweight Convolutional Neural Network Architectures for Real-Time Surveillance

Rani and Kumar [26] developed the MobileNet model for human activity recognition in smart surveillance. The lightweight Convolutional Neural Network (CNN) architecture was well-suited for real-time and resource constrained applications such as human activity recognition. The MobileNet model was trained on a large dataset that replaced the final layer with a task-specific

layer (dense layers or classification heads) to recognize human activities. MobileNet's small size and computational efficiency made it ideal for devices with limited resources, such as smartphones. However, MobileNet struggled to accurately recognize complex or subtle human activities, particularly when the dataset was limited.

D. Transformer-Based and Graph-Oriented Frameworks

Shin *et al.* [27] implemented a multistage DL approach to detect abnormal behaviors. Initially, features were extracted using a ViT-based CLIP model and a CNN-based I3D module enhanced with Temporal Contextual Aggregation (TCA). These features were concatenated and processed by Uncertainty-Regulated Dual Memory Units (UR-DMU), which integrated GCN and Multi-Head Self Attention mechanisms to capture video-level associations. Although comprehensive, this approach required significant computational resources and posed scalability challenges.

E. Attention-Based Spatio-Temporal Models

Ullah *et al.* [28] presented a deep learning-based model that employed a Deep CNN (DCNN) for spatial feature extraction and Temporal Convolutional Networks (TCN) with multi-head attention to model the temporal dependencies. This structure improved the ability of the approach to detect anomalies in complex environments. However, the system occasionally struggled with generalization across varied surveillance conditions involving occlusion or overlapping actions.

F. Research Gap

Despite notable advancements in video anomaly detection, the existing models continue to face challenges in effectively balancing spatial and temporal information. Many prior approaches disproportionately emphasize their spatial feature extraction or temporal modeling, without fully leveraging their synergy. Moreover, several deep learning architectures exhibit high computational overhead or limited adaptability to real-world noisy datasets. Most importantly, the integration of dynamic attention mechanisms with bidirectional temporal modeling remains underexplored. This study addresses these issues through an optimized CNN-BiLSTM-STAT framework equipped with spatio-temporal adaptive modules. An important innovation of this research involves the CNN-BiLSTM-Transformer integration present in the synergistic incorporation of different complementary architectures through the specialized Bidirectional Temporal Difference with STAT mechanism. Whereas CNNs extract multi-scale spatial features and BiLSTMs model the bidirectional temporal dependencies, the Transformer constituent with the present CNN-BiLSTM module offers adaptive attention, which dynamically balances motion features and static semantics in terms of scene conditions. This three-way incorporation specifically solves the limitations of existing methods through: (1) modeling both forward and backward temporal contexts simultaneously, (2) adaptively weighting spatial features based on their relevance to anomaly detection, and (3) utilizing the global background with multi-head attention while maintaining computational efficiency. The proposed approach introduces bidirectional temporal variation analysis and spatial adaptive attention, which are particularly developed for video anomaly detection, enabling the model to focus on critical regions and motion patterns that classify the anomalies from normal activities. This aimed incorporation of spatial, temporal, and adaptive attention mechanisms illustrates an important development across generic integrations of these architectures.

III. Proposed Methodology

In the proposed approach for video-based anomaly detection, this study leveraged a sophisticated combination of existing approaches to enhance the accuracy and robustness of anomaly identification. Most existing attention-based methods in video anomaly detection apply static attention or concentrate only on either spatial or temporal information. On the other hand, the proposed CNN-BiLSTM-STAT module provides an integrated and adaptive spatio-temporal attention mechanism. The novelty of the proposed method lies in a motion-aware spatio-temporal adaptive attention mechanism that explicitly models bidirectional temporal differences and dynamically fuses motion and appearance features, rather than simply combining CNN, BiLSTM, and transformer components. It explicitly estimates bidirectional temporal variations among consecutive frames to emphasize the motion changes that are important for detecting anomalous activities. Moreover, an adaptive fusion mechanism dynamically balances motion features and appearance features based on the anomaly type. Moreover, spatial attention is directed through temporal motion differences, enabling the model to concentrate on regions exhibiting abnormal motion rather than uniformly attending to salient objects. This adaptive and motion-guided attention design distinguishes the proposed approach from existing fixed-attention models.

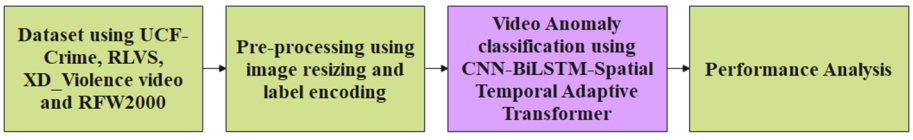

Fig. 1. Workflow of the proposed method.

Fig. 1 shows the workflow of the proposed model, which uses a multistage DL approach for anomaly detection in video surveillance.

A. Dataset

Anomaly detection dataset has become an important

part of developing and estimating approaches aimed at determining irregular or unexpected events within data streams. These datasets provide diverse scenarios, enabling researchers to train and test their models under various constraints. In addition, there are diverse dataset available for anomaly detection, which utilizes subsequent benchmark dataset such University of Central Florida Crime (UCF-Crime) dataset [29], Real-Life Violence Situations (RLVS) dataset [30], Extreme Dataset for Violence Detection (XD_Violence) [31], and Real-World Fight 2000 (RFW-2000) dataset [32], which offer different scales, backgrounds, and types of anomalies catering to different research needs. By utilizing these dataset researchers can benchmark their anomaly detection methods, assess their performance, and contribute to advancing the field of anomaly detection in real-world applications.

1) UCF-crime dataset

The UCF-Crime dataset consists of a large-scale anomaly detection dataset, which includes 1900 untrimmed videos collected from real-world streets and indoor surveillance cameras with a total duration of 128 h. The dataset covers 13 real-world anomalies such as arrest, arson, accident, burglary, stealing, explosion, abuse, fighting, shooting, assault, vandalism, shoplifting, and robbery. Compared with the static background in the ShanghaiTech dataset, the UCF-Crime [8] dataset contains the most complex and diverse backgrounds. The training set of the UCF-Crime dataset contained 1610 videos with 800 labelled as normal and 810 labelled as anomalous. The testing set involved 290 videos, 150 labelled as normal and 140 labelled as anomalous, and included frame-level labels.

2) RLVS dataset

The RLVS dataset contains 2000 video clips, which are equally partitioned among violent and everyday activities. Violent videos depict physical arguments in diverse settings such as streets, prisons, and schools. The videos in this dataset involve features with high resolutions in the range of 480 p and 720 p in a spanning duration of 3–7 s. The frame extraction process was performed using a 10-frame interval, resulting in six frames per second. Notably, some frames within violent and shoplifting videos were eliminated during the data-cleaning phase because they did not depict the relevant actions and were more similar to frames from normal videos.

3) XD-violence video dataset

The XD-Violence dataset comprises a variety of media formats, specifically videos and audio. The dataset encompasses a diverse range of backgrounds such as movies, games, and live scenes. It consists of 4754 videos, with 3954 videos designated for training purposes and equipped with video-level labels. Additionally, 800 testing videos were labeled at the frame level.

4) RWF-2000 dataset

This dataset consists of 2000 real-world surveillance videos of normal and violent activities collected from publicly available YouTube sources. Each video is approximately 5 s in duration, with varying resolutions and a frame rate of 30 Frames Per Second (FPS). The dataset includes different types of violent activities involving two-person, crowd, and individual-person aggressive actions.

B. Pre-processing

Preprocessing is a significant stage in ensuring the excellence and reliability of a dataset for accurate classification results. Different preprocessing techniques, such as image resizing and label encoding, are employed to improve image quality and prepare data for classification tasks.

1) Image resizing

In the preprocessing stage, each video frame underwent standardization to ensure reliability over the dataset. The resizing operation adjusts the dimensions of the input frames, reducing the computational load and unifying the input shapes for model compatibility. In addition, the frames are converted into an 8-bit format, which simplifies memory management and ensures uniform image quality across all inputs [33].

2) Label encoding

Label encoding was used to prepare the categorical data for model training. Because most deep-learning models require numerical inputs, categorical class labels were transformed accordingly. Among the various encoding techniques, one-hot encoding was selected owing to its simplicity and effectiveness. This method assigns a unique binary vector to each class, enabling the model to interpret the data without introducing ordinal relationships [34]. All videos were retained at their original frame rates without temporal resampling. For each video clip, 10 frames were uniformly sampled across the entire duration to ensure temporal coverage, irrespective of the original frames per second. The extracted frames were resized to 64×64×3 and normalized before being fed into the network. Rather than sampling only 10 frames from an entire video, the proposed method performs a non-overlapping temporal windowing strategy. Every video is segmented into consecutive windows of 10 frames, and feature extraction and classification are performed at the window level. This design makes sure that short and sudden anomalous events are modelled within at least one window whereas preserving temporal motion continuity. The use of fixed-size windows also minimizes computational overhead and supports efficient training.

C. Video Anomaly Classification Using CNN-BiLSTM-STAT

Video anomaly classification requires the extraction of both spatial and temporal features from video sequences. To address this, the proposed model integrates a CNN, BiLSTM units, and transformer modules. CNNs were used to extract meaningful spatial patterns from individual frames because they performed well with grid-structured data. However, CNNs lack the capability to capture temporal dependencies between frames, which is essential for understanding motion and context in video sequences. To overcome this, BiLSTM networks were incorporated to process frame sequences in both forward and backward directions. This bidirectional processing enables an

approach to learn dependencies over time more effectively. Despite their strengths in temporal modeling, BiLSTMs present challenges in terms of computational overhead and inefficiency when dealing with longer input sequences.

Transformer modules were introduced into the architecture to address this bottleneck. Their multi-head attention mechanism allowed for global context modeling and supports parallel processing, thereby enhancing the model's scalability and performance. The Transformer component complements the BiLSTM and CNN by enabling a deeper understanding of the temporal structure and improving sequence-level prediction accuracy. Together, the CNN, BiLSTM, and Transformer components form a hybrid architecture, termed CNN-BiLSTM-STAT. The STAT varies from a conventional Transformer by jointly modelling temporal attention and spatial adaptive weighting. Temporal self-attention is directed through the bidirectional temporal variations, whereas spatial adaptive attention highlights motion-salient regions within frames. This design enables efficient global dependency modeling without relying on heavy transformer architectures. This configuration allows for robust spatio-temporal feature representation. For further optimization, a grid search and cross-validation were employed to fine-tune the hyperparameters. The architecture uses different parallel CNN divisions: one with a smaller convolutional kernel and the other with a larger kernel to capture features at multiple scales. After feature fusion, the data were passed through the BiLSTM layers to model temporal dynamics. Subsequently, the Transformer encoder with multi-head attention processes the outputs to integrate the global sequence information. Finally, a fully connected layer is used for multiclass classification of anomalies. Fig. 2 illustrates the architecture of the proposed method.

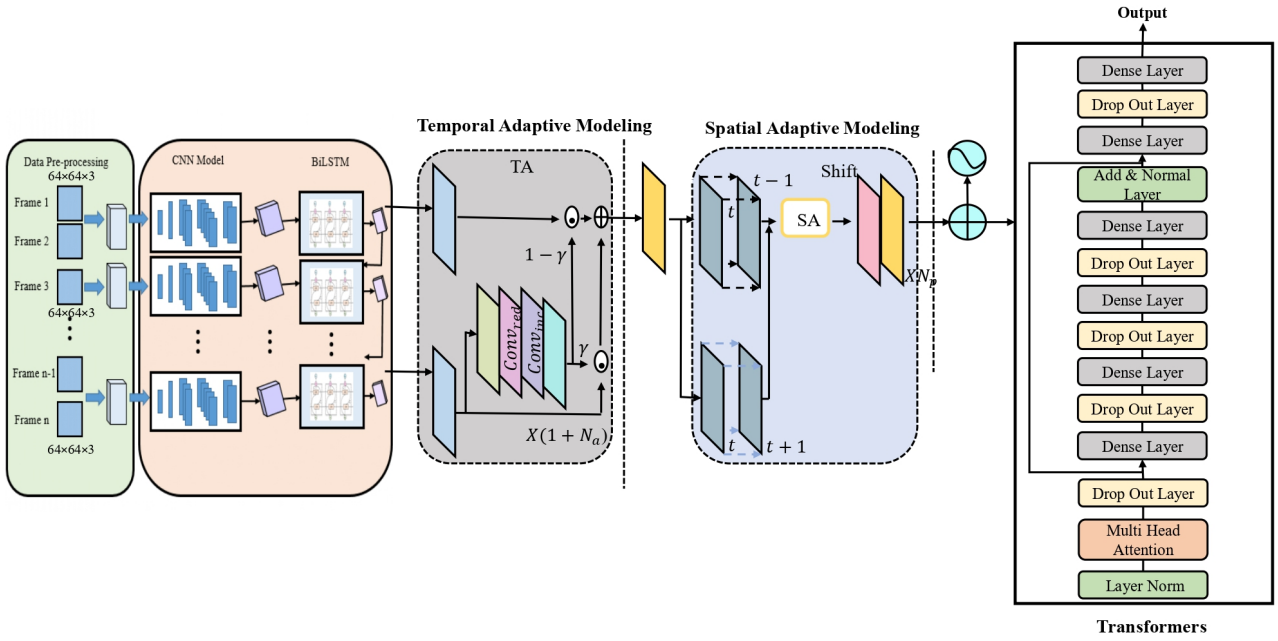

Fig. 2. Architecture of the proposed method.

1) Convolutional neural network

CNN demonstrates a strong ability for the automatic extraction of meaningful features from input data by modelling the underlying spatial patterns. Its core operation, convolution, involves performing element-wise multiplication and summation over the sliding windows across the input. In the proposed approach, different parallel convolutional subdivisions are utilized: one applies a smaller kernel of size 3, whereas the other utilizes a larger kernel of size 5 to extract multi-scale spatial features. Every subdivision performs the input using a 1D Convolutional layer (Conv1d) for initial feature extraction, followed by an activation function that introduces nonlinearity. This is achieved by MaxPooling (MaxPool1d) to down sample the feature maps, Batch Normalization (BatchNorm1d) to enhance training stability, convergence and a flattening operation to convert the multidimensional output into a one-dimensional feature vector. Finally, the outputs from both branches were concatenated to design a unified feature depiction for subsequent processing. A convolution operation is mathematically formulated in Eq. (1).

$$y(t) = (x \times w)(t) = \sum_{i=0}^{k-1} x(t+i)w(i) \tag{1}$$

where $x(t)$ denotes the input image, $w(i)$ denotes the convolutional kernel, $y(t)$ denotes the output data, and k denotes the convolutional kernel size.

Although CNN significantly extracts the local features, it fails to sufficiently analyze the temporal correlations in the data. This constraint removes the CNN from completely modelling the temporal features in the video anomaly data. To solve the problem of capturing temporal patterns in anomalous video data, the model integrates a BiLSTM network. BiLSTM leverages different LSTM units to function as input sequences in both directions,

allowing for a more comprehensive understanding of the temporal dependencies.

2) BiLSTM

The LSTM architecture effectively addresses the long-term dependency challenges typically faced by traditional RNNs using different gate inputs i_t, forgets, f_t and output o_t. These gating mechanisms enable LSTM to retain and proliferate crucial data across extended sequences, thereby reducing the risk of vanishing gradients and enhancing learning stability. In BiLSTM, both the forward and backward outputs are combined to create a more comprehensive feature depiction, which is subsequently passed to the following layers in the model. Internal gating mechanisms play a crucial role in managing the data flow. The input gate regulates the amount of the current input that contributes to updating the memory state. A forget gate identifies the extent to which previous memory information is retained or discarded at the current time step. These gates work together to preserve the relevant context while filtering out less important data, enabling an approach to effectively model long-term dependencies in sequential input. The output gate identifies the contribution of the present memory data to the output. A current memory cell c_t is developed through the weighted hybridization of the past memory cell c_{t-1} and the present input data. A hidden state h_t is regulated through the output gate and combined with the processed data from the memory cell. The mathematical expression of LSTM gates is formulated in Eqs. (2)–(6).

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tan h\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{5}$$

$$h_t = o_t \tan h\left(c_t\right) \tag{6}$$

where W_{xi}, W_{xf}, $W_{xo} \in R^{d\times h}$, and W_{hi}, W_{hf}, $W_{ho} \in R^{h\times h}$ are the weight parameters, b_i, b_f and b_o are the deviation parameters. The candidate memory cell $\tilde{C}_t$ was then estimated. It utilizes the $tanh$ activation function in the range between [−1, 1] as the activation function.

3) Bidirectional variation based optimal temporal adaptive module

This module was designed to learn critical motion-related features and adaptively integrate both dynamic and static appearance information. While bidirectional temporal difference captures sufficient motion cues across frames, it remains essential to effectively embed this motion information into a static semantic representation to enhance contextual understanding. To address this issue, an adaptive fusion strategy is introduced. This strategy dynamically adjusts the fusion weights between the motion features and static semantics based on varying scene conditions, allowing the model to emphasize the most relevant aspects of the input for accurate anomaly detection. Assume $D = \frac{1}{2Conv_1(D_f)} + \frac{1}{2Conv_2(D_b)}$, where $Conv_1$ and $Conv_2$ are architecturally reliable through static features. This research utilizes the adaptive fusion module to understand the influence of dynamic motion, as formulated in Eq. (7), as follows:

$$\gamma = Sigmoid\left(Conv_{enc}\left(Conv_{\min}\left(Avgpool\left(D\right)\right)\right)\right) \tag{7}$$

Eq. (7) illustrates the process of enhancing the utilization of global spatial information. Initially, the spatial details are summarized through average pooling, which reduces the spatial dimension of D to 1. Subsequently, a 1×1 2D channel-minimized Convolution $Conv_{min}$ is applied to own-sample a channel through a factor of τ, pursued through the 1×1 2D channel-enhanced Convolution $Conv_{enc}$ to obtain the channel dimension. Eventually, the output is constrained between 0 and 1 after the sigmoid activation. Therefore, the causes of dynamic motion and static semantics are balanced by fusion computation as formulated in Eq. (8):

$$X' = \left(1-\gamma\right) \times Conv_3\left(X\right) + \gamma \times D \tag{8}$$

where, $Conv_1$, $Conv_2$ and $Conv_3$ share a similar structure with various parameters, primarily a 7×7 convolution layer. Different temporal adaptive layers are applied to perform the combined features and eventually acquire a temporal motion that improves depiction $\hat{X}$. The temporal adaptive module enables effective temporal modeling by leveraging bidirectional temporal variations along with the multiple learnable convolutional layers. This design allows the network to accurately capture motion dynamics and precisely localize moving objects within the video sequences.

4) Bidirectional variation based optimal spatial adaptive module

Although temporal adaptive modelling enhances the encoding of temporal information across frames in the lower layers of the network, it does not adequately capture the critical spatial features within individual frames. Moreover, variability in spatial information across frames can obscure the essential semantic cues required for accurate action recognition. To address this, the proposed approach employs a bidirectional-variation-based optimal spatial adaptive module. This module emphasizes both semantically relevant and position-sensitive features by leveraging motion cues in the higher layers of the network, thereby improving the recognition performance. After computing the bidirectional temporal difference, the model applies a 1×1 2D convolution followed by an activation function to construct a spatial attention map that

highlights key spatial regions for further processing as formulated in Eq. (9) as follows:

$$A_s = Sigmoid\left(Conv_{\min}\left(D\right)\right) - \delta_s \tag{9}$$

where the channel-minimized convolution $Conv_{min}$ is a 1×1 convolution that is utilized to combine the channel data to generate a preliminary spatial attention map. The sigmoid function is utilized for normalizing the learned attention coefficients. δ_s serves as a hyperparameter that controls the intensity of the spatial-wise attention map.

Algorithm 1 shows the pseudocode of the proposed method for improved reproducibility.

Algorithm 1. Pseudocode of the proposed method

Pseudocode: Proposed CNN-BiLSTM-STAT approach for video anomaly detection
START
Input: VideoClip V with 10 frames of size 64×64×3
Output: Predicted Class Label
Step 1: Preprocessing
For each video clip V:
Extract 10 frames
Resize each frame to 64×64×3
Apply label encoding to class label
Step 2: Spatial Feature Extraction using CNN
For each frame *Fi* in V:
Fi = *Conv2D*(32 *filters*, 3×3, *ReLU*) → *MaxPooling*(2×2)
Fi = *Conv2D*(64 *filters*, 3×3, *ReLU*) → *MaxPooling*(2×2)
Fi = *Flatten* (*Fi*)
Output: Frame Features = [F1, F2, ..., F10] → Shape (10, feature_dim).
Step 3: Temporal Modeling using Bidirectional LSTM
Input: Frame Features
Temporal Features = BiLSTM (units = 64, return_sequences = True)
Output: Temporal Features → Shape: (10, 128)
Step 4: Temporal Adaptive (TA) Module
Df = *Conv*(*ForwardDifference*(*TemporalFeatures*))
Db = *Conv*(*BackwardDifference*(*TemporalFeatures*))
D = 0.5×*Df* + 0.5×*Db*
γ = *Sigmoid*(*Conv*(*AvgPool*(*D*)))
TA_Output = (1 − γ)×*Conv*(*TemporalFeatures*) + γ×*D*
Step 5: Spatio–Temporal Adaptive Transformer (STAT)
Input: TA_Output
Compute temporal self-attention using bidirectional temporal variation
Apply spatial adaptive attention guided by temporal differences
Fuse spatial and temporal attention features
Apply feedforward layer with residual connection and layer normalization
Output: STAT-Encoded Features
Step 6: Spatial Adaptive (SA) Module
For each timestep t in Residual_Encoded:
Df = Conv($\hat{X}$ [*t*]) − $\hat{X}$ [*t* − 1]
Db = *Conv*($\hat{X}$)[*t*] − $\hat{X}$ [*t*+1]
D = *Df* + *Db*
As = *Sigmoid*(*Conv*(*D*)) − δs
Ac = *Sigmoid*(*Conv*(*AvgPool*(*D*))) − δc
SA _ Output[*t*] = $\hat{X}$[*t*] + $\hat{X}$[*t*] ⊙ *As* + $\hat{X}$[*t*] ⊙ *Ac*
Step 7: Global Feature Pooling
Global Features = Average (SA_Output over time axis)
Step 8: Classification
Global Features = Dense (64, ReLU) → Dropout (0.3)
Output = Dense (num_classes, Softmax)
Return: Predicted class label from Output
END

To ensure that the attention-enhanced features remain influenced primarily by an input, the model combines the original and attention-weighted features through element-wise multiplication and addition. This operation selectively amplifies the key spatial features while preserving essential background information, thereby allowing the network to concentrate on both the position-relevant and semantically significant regions. The Spatial Adaptive (SA) module generates a distinct attention map for every frame by utilizing the bidirectional temporal differences, ultimately enhancing the representational capacity of the model and improving the precision of action recognition.

IV. Experimental Results

This section evaluates the proposed anomaly recognition model by using the four publicly available benchmark dataset that are widely adopted for video anomaly detection tasks.

The robustness and effectiveness of the model are demonstrated through the comprehensive experimental results. The findings confirm that the integration of spatial and temporal attention mechanisms crucially improves the model's capability to accurately identify the anomalous events. The significance of the proposed approach was implemented on MATLAB R2020b with system configurations of 16 GB RAM, Intel i5 processor, Windows 10 OS, and 6 GB GPU. The training and testing of the datasets are partitioned as 80% and 20% individua ally. The dataset was partitioned at the video level, such that entire video clips are assigned exclusively to either the training or testing set. No frames extracted from a provided video appear in both partitions. Frame extraction, resizing, and normalization were performed independently after the train-test split to prevent any information leakage between the two sets. The effectiveness of the proposed approach is estimated through various performance metrics such as accuracy, precision, recall, specificity, and F1-Score. The mathematical expressions of these performance metrics are formulated in Eqs. (10)–(14) as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{14}$$

where *TP* is the True Positive, *TN* is the True Negative, *FP*

is the False Positive, and *FN* is the False Negative.

Table I illustrates the hyperparameter settings of the proposed method.

TABLE I. HYPERPARAMETER SETTINGS OF THE PROPOSED METHOD

Parameters	Values
Activation function	Softmax
Loss function	Binary cross entropy for RLVS and RWF, Categorical Cross entropy for UCF-Crime and XD-Voilence
Optimizer	Adam
Learning Rate	0.00001
Batch size	32
Epoch	18

A. Performance Analysis

The effectiveness of the proposed approach was evaluated by using the four different datasets. To estimate the classification results, the proposed approach was compared with different classifiers such as CNN-LSTM, Bi-LSTM, and CNN. Table II represents the comparative performance of various classification models across the four benchmark datasets using metrics such as accuracy, precision, recall, F1-Score, and AUC. The proposed CNN-BiLSTM-STAT approach significantly outperforms the existing models such as YOLO, CNN+GRU+Transformer, 3DCNN, Temporal CNN (T-CNN), Swin Transformer, and TimeSformer. Notably, it achieved near-perfect performance on the UCF-Crime dataset with 99.91% accuracy and 100% AUC by demonstrating its robustness and superiority in anomaly classification tasks.

TABLE II. PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFICATION RESULTS WITH TRANSFORMER

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
YOLO	UCF-Crime	96.75	95.42	94.58	94.93	95.83
	RLVS	88.21	87.53	87.02	87.24	87.9
	XD_Violence_video	84.37	82.11	83.81	82.97	88.14
	RWF2000	90.28	89.26	89.54	89.39	90.32
CNN + GRU + Transformer	UCF-Crime	97.32	96.81	95.67	96.23	97.01
	RLVS	89.67	88.93	88.45	88.69	88.74
	XD_Violence_video	86.79	85.62	86.41	86.01	89.23
	RWF2000	91.36	90.78	91.05	90.91	91.84
3DCNN	UCF-Crime	95.84	94.31	93.22	93.76	94.89
	RLV	87.19	86.34	85.93	86.13	86.77
	XD_Violence_video	83.66	81.74	82.9	82.31	87.21
	RWF2000	89.58	88.47	88.72	88.59	89.63
T-CNN	UCF-Crime	95.23	94.07	93.45	93.76	94.12
	RLVS	86.48	85.91	85.36	85.63	86.02
	XD_Violence_video	82.90	81.23	82.47	81.84	86.35
	RWF2000	88.67	87.92	88.1	88.01	88.93
Swin Transformer	UCF-Crime	97.84	97.32	96.88	97.1	98.02
	RLVS	90.25	89.83	89.6	89.71	90.08
	XD_Violence_video	86.47	84.95	86.21	85.57	89.76
	RWF2000	91.92	91.47	91.3	91.38	92.6
TimeSformer	UCF-Crime	98.29	97.83	97.45	97.64	98.39
	RLVS	91.12	90.7	90.32	90.51	91.24
	XD_Violence_video	87.74	86.18	87.42	86.79	90.33
	RWF2000	92.45	92.18	92.1	92.14	93.02
CNN-BiLSTM-STAT	UCF-Crime	99.91	99.88	99.82	99.85	100
	RLVS	93.00	93.02	93.00	93.00	92.67
	XD_Violence_video	89.68	87.55	90.89	88.70	93.10
	RWF2000	93.75	93.84	93.75	93.75	94.50

Table III details the class-wise evaluation of the proposed model over the UCF-Crime, RLVS, XD_Violence, and RWF-2000 datasets. Each anomaly class, such as assault, vandalism, robbery, and abuse, is individually assessed in terms of accuracy, precision, recall, and F1-Score. The results show consistently high scores across all metrics, with some classes, such as abuse and fighting, achieving perfect scores, which confirms the model's precise and effective anomaly localization and classification capabilities. Particular anomaly classes exhibit 100% precision because no false positive predictions were observed for those categories in the test set. This behavior is significantly present in well-separated anomaly classes with highly distinctive motion patterns. However, such results are dataset-dependent and do not imply universal perfect classification.

Table IV represents a performance analysis of the computational costs of the proposed method. This table compares the computational efficiency of the proposed method with other transformer-based models using key indicators such as training time, memory usage, inference time, FLOPs, and parameter count. The proposed model exhibited lower training time and memory consumption while maintaining high accuracy. For instance, on the UCF-Crime dataset, it recorded the lowest training time (586.26 s) and inference time (1.55 s) while reflecting its computational efficiency for real-time deployment.

Table V represents the statistical evaluation of the proposed method. Statistical analysis involved *p*-values, confidence intervals and accuracy for each model and dataset. The proposed model demonstrates a statistically significant improvement over others with the lowest *p*-

values (e.g., 0.015 for UCF-Crime) and the narrowest confidence intervals, indicating highly reliable performance. The UCF-Crime dataset shows an outstanding accuracy of 99.91%, validating the consistency and precision of the model in anomaly detection. The statistical significance is evaluated using a paired t-test to compare the proposed CNN-BiLSTM-STAT model with competing methods. The null hypothesis (H_0) assumes that no statistically significant difference exists between the compared methods. The alternative hypothesis (H_1) assumes a significant performance difference. The sample size corresponds to the number of test video samples in each dataset; a significance level of $\alpha = 0.05$ was adopted. The sample size corresponds to the number of test video samples in each dataset.

TABLE III. CLASS WISE RESULTS OF THE PROPOSED METHOD USING ALL DATASET

Dataset	Class	Precision (%)	Recall (%)	F1-Score (%)
UCF-Crime	RoadAccidents	100	99.17	99.59
	Assault	100	99.81	99.91
	Vandalism	100	99.82	99.91
	Arrest	99.91	100	99.96
	Shooting	99.99	99.71	99.85
	NormalVideos	99.92	99.99	99.96
	Arson	98.62	99.68	99.15
	Explosion	99.95	99.71	99.83
	Shoplifting	100	99.99	99.99
	Robbery	100	99.76	99.88
	Stealing	99.90	99.90	99.90
	Burglary	100	99.90	99.95
	Abuse	100	100	100
	Fighting	100	100	100
	Average	99.88	99.82	99.85
RLV	Non violence	93.88	92.00	92.93
	Violence	92.16	94.00	93.07
	Average	93.02	93.00	93.00
XD_Violence_video	Fighting	73.20	93.33	82.05
	Normal	98.87	87.67	92.93
	Shooting	90.59	91.67	91.12
	Average	87.55	90.89	88.70
RWF2000	Fight	91.87	96.00	93.89
	NonFight	95.81	91.50	93.61
	Average	93.84	93.75	93.75

TABLE IV. PERFORMANCE ANALYSIS OF COMPUTATIONAL COST OF THE PROPOSED METHOD

Methods	Dataset	Training time per epoch (s)	Memory usage (MB)	Inference time (s)	Flops	Parameters (Million)
YOLO	UCF-Crime	732.11	14,872.33	1.76	94.73	37.84
	RLV	69.44	5220.31	0.58		
	XD_Violence_video	84.62	6705.21	0.62		
	RWF2000	56.12	6112.93	0.51		
CNN + GRU + Transformer	UCF-Crime	653.38	13,521.11	1.64	87.15	34.29
	RLV	62.18	4950.29	0.56		
	XD_Violence_video	77.12	6307.1	0.61		
	RWF2000	49.88	5824.66	0.48		
3DCNN	UCF-Crime	690.77	13,950.88	1.69	91.62	36.17
	RLV	65.55	5099.34	0.57		
	XD_Violence_video	82.34	6555.22	0.63		
	RWF2000	52.01	5960.48	0.49		
Temporal CNN (T-CNN)	UCF-Crime	705.2	14,120.17	1.72	92.45	32.89
	RLV	67.88	5150.07	0.57		
	XD_Violence_video	79.25	6438.66	0.62		
	RWF2000	53.66	6011.34	0.50		
Proposed CNN-BiLSTM-STAT	UCF-Crime	586.26	13,647.30	1.5526	85.93	32.34
	RLV	47.23	5176.79	0.4633		
	XD_Violence_video	66.45	6108.02	0.4938		
	RWF2000	38.21	6044.51	0.3978		

Table VI analyzes the generalizability of the model by training on the XD_Violence dataset and testing on RLVS and RWF2000. The proposed CNN-BiLSTM-STAT model again delivers superior results across all metrics compared to the other methods, achieving 89.09% accuracy and 88.29% F1-Score on RWF2000. This underscores the adaptability and transfer-learning capability of the model across various video anomaly datasets.

Table VII illustrates the ablation study conducted in this research. This ablation study evaluated the contribution of individual components such as CNN, Bi-LSTM, and the

STAT module. The ablation findings emphasize the contribution of discrete architectural components. The conventional CNN provides robust spatial representations from video frames, but ignores temporal context, leading to constrained detection effectiveness. Integrating the BiLSTM resulted in a significant enhancement through modeling bidirectional temporal dependencies over the frames. An integration of STAT attains the most significant performance gain, as it dynamically incorporates the motion-aware temporal attention with spatial adaptation. This progressive improvement confirms that every component offers meaningfully and that the combined architecture is important for robust anomaly detection.

TABLE V. STATISTICAL ANALYSIS OF THE PROPOSED METHOD

Methods	Dataset	*P* value	Confidence interval	CI (±)	Accuracy (%)
YOLO	UCF-Crime	0.031	[95.82–97.68]	±0.93	96.75
	RLVS	0.044	[87.31–89.11]	±0.90	88.21
	XD_Violence_video	0.052	[83.28–85.46]	±1.09	84.37
	RWF2000	0.039	[89.24–91.32]	±1.04	90.28
CNN + GRU + Transformer	UCF-Crime	0.028	[96.59–98.05]	±0.73	97.32
	RLVS	0.037	[88.65–90.69]	±1.02	89.67
	XD_Violence_video	0.042	[85.61–87.97]	±1.18	86.79
	RWF2000	0.033	[90.19–92.53]	±1.17	91.36
3DCNN	UCF-Crime	0.051	[94.45–97.23]	±1.39	95.84
	RLVS	0.056	[86.06–88.32]	±1.13	87.19
	XD_Violence_video	0.064	[82.09–85.23]	±1.57	83.66
	RWF2000	0.027	[88.32–90.84]	±1.26	89.58
T-CNN (Temporal CNN)	UCF-Crime	0.023	[94.10–96.36]	±1.13	95.23
	RLVS	0.021	[85.42–87.54]	±1.06	86.48
	XD_Violence_video	0.027	[81.56–84.24]	±1.34	82.90
	RWF2000	0.039	[87.41–89.93]	±1.26	88.67
CNN-BiLSTM-STAT	UCF-Crime	0.015	[99.43–100.39]	±0.48	99.91
	RLVS	0.018	[92.17–93.83]	±0.83	93.00
	XD_Violence_video	0.011	[88.22–91.14]	±1.46	89.68
	RWF2000	0.016	[92.71–94.79]	±1.04	93.75

TABLE VI. CROSS-DATASET ANALYSIS WHICH THE MODEL IS TRAINED ON XD_VIOLENCE_VIDEO TESTED ON RLVS DATASET AND RWF2000

Method	Datatset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)
MViT (Multiscale Vision Transformer)	RLVS	83.17	82.49	81.92	82.26	84.53
	RWF2000	86.43	85.66	85.24	85.37	87.39
CNN + GRU + Transformer	RLVS	84.92	84.27	83.74	83.98	85.86
	RWF2000	87.21	86.65	86.18	86.41	88.19
3DCNN	RLVS	81.72	81.08	80.41	80.68	83.27
	RWF2000	85.11	84.33	83.92	84.11	86.18
T-CNN (Temporal CNN)	RLVS	80.58	79.97	79.44	79.7	81.91
	RWF2000	84.18	83.58	83.23	83.4	85.42
CNN-BiLSTM-STAT	RLVS	86.69	86.14	85.78	85.95	87.83
	RWF2000	89.09	88.42	88.17	88.29	90.03

TABLE VII. ABLATION STUDY OF THIS RESEARCH

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
CNN	UCF-Crime	94.52	93.74	92.8	93.26	94.05
	RLVS	85.76	84.9	84.45	84.67	85.42
	XD_Violence_video	81.38	80.12	80.87	80.49	85.03
	RWF2000	87.92	87.13	87.25	87.19	88.07
Bi-LSTM	UCF-Crime	95.13	94.58	93.91	94.24	95.22
	RLVS	86.92	86.01	85.47	85.74	86.63
	XD_Violence_video	82.47	81.26	82.09	81.67	86.12
	RWF2000	88.76	88.11	88.22	88.16	89.12
CNN-Bi-LSTM	UCF-Crime	94.52	93.74	92.8	93.26	94.05
	RLV	85.76	84.90	84.45	84.67	85.42
	XD_Violence_video	81.38	80.12	80.87	80.49	85.03
	RWF2000	87.92	87.13	87.25	87.19	88.07
CNN with STAT Module	UCF-Crime	98.41	97.96	97.58	97.76	98.64
	RLV	90.88	90.67	90.53	90.6	90.21
	XD_Violence_video	87.36	85.41	87.07	86.23	90.44
	RWF2000	91.58	91.43	91.36	91.39	92.33
CNN-BiLSTM-STAT	UCF-Crime	99.91	99.88	99.82	99.85	100
	RLV	93.00	93.02	93.00	93.00	92.67
	XD_Violence_video	89.68	87.55	90.89	88.70	93.10
	RWF2000	93.75	93.84	93.75	93.75	94.50

The results show a significant performance gain while combining CNN and Bi-LSTM, which was further improved by integrating the STAT module and transformer. The complete model achieved the highest scores across all the datasets (e.g., 99.91% accuracy on UCF-Crime), by proving the necessity and effectiveness of the proposed architecture.

Fig. 3 shows the qualitative visualization of anomaly localization using Grad-CAM. To enhance interpretability, Grad-CAM visualizations were incorporated to illustrate how the proposed model focuses on critical regions of video frames while detecting the anomalies. These visual explanations provide evidence that the CNN-BiLSTM-STAT framework does not rely on spurious correlations but attends to semantically relevant features such as aggressive actions or accident regions.

Fig. 4 illustrates the accuracy of the proposed method across all four datasets (UCF-Crime, RLVS, XD_Violence, and RWF-2000). Each subfigure (a–d) corresponds to a dataset showing high accuracy consistently above 89%, with UCF-Crime near 100%.

Fig. 3. Qualitative visualization of anomaly localization using Grad-CAM.

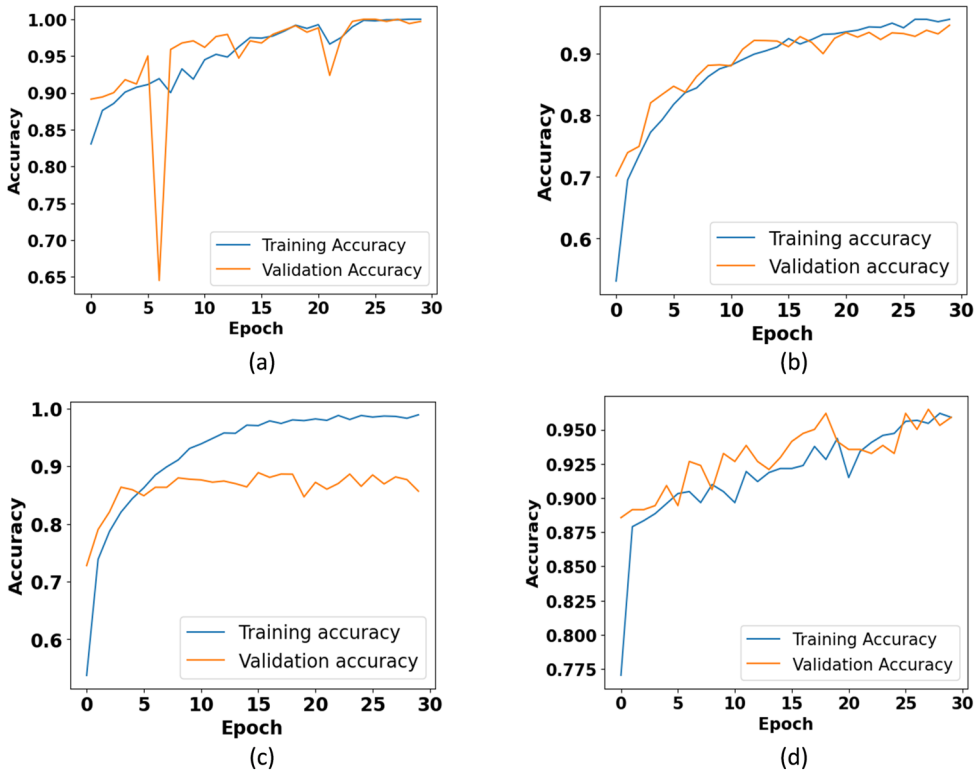

Fig. 4. Visual representation of accuracy results of the proposed method: (a). UCF-Crime dataset, (b). RLVS dataset, (c). XD_Violence video dataset, (d). RFW-2000 dataset.

Fig. 5 shows the training loss curves for the same four datasets. This demonstrates that the proposed model achieves rapid convergence with minimal overfitting, as evidenced by the sharp drop and stabilization of the loss. This visual trend suggests effective learning and generalization during the training.

Fig. 6 shows the confusion matrices for each dataset, providing insight into classification reliability. The matrices exhibited dense diagonals and minimal misclassifications, indicating the strong ability of the model to correctly classify each class. For example, in UCF-Crime, nearly all predicted labels align with the ground truth values.

Fig. 7 displays the ROC curves for the four datasets, illustrating the trade-off between the true positive rate and false positive rate. The proposed method achieved AUC values close to 1.0 across all the dataset with UCF-Crime reaching the ideal AUC of 1.0. These results confirm the superior discriminative capability of the model in distinguishing anomalies from normal events.

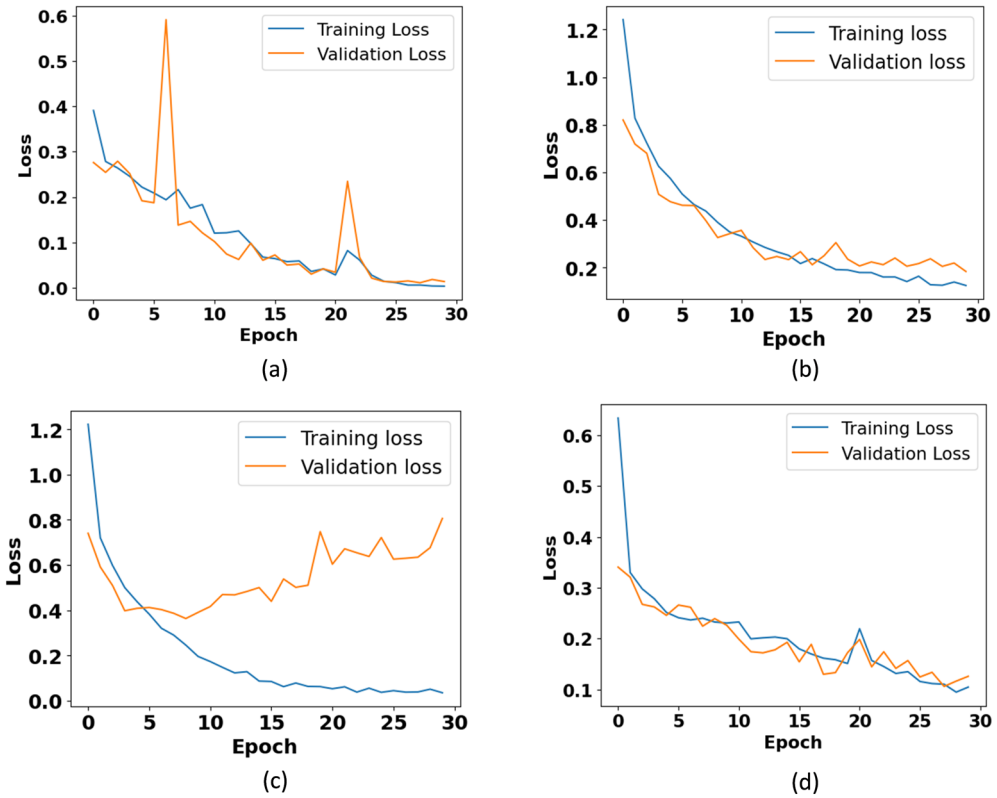

Fig. 5. Visual representation of loss results of the proposed method: (a). UCF-Crime dataset, (b). RLVS dataset, (c). XD_Violence video dataset, (d). RFW-2000 dataset.

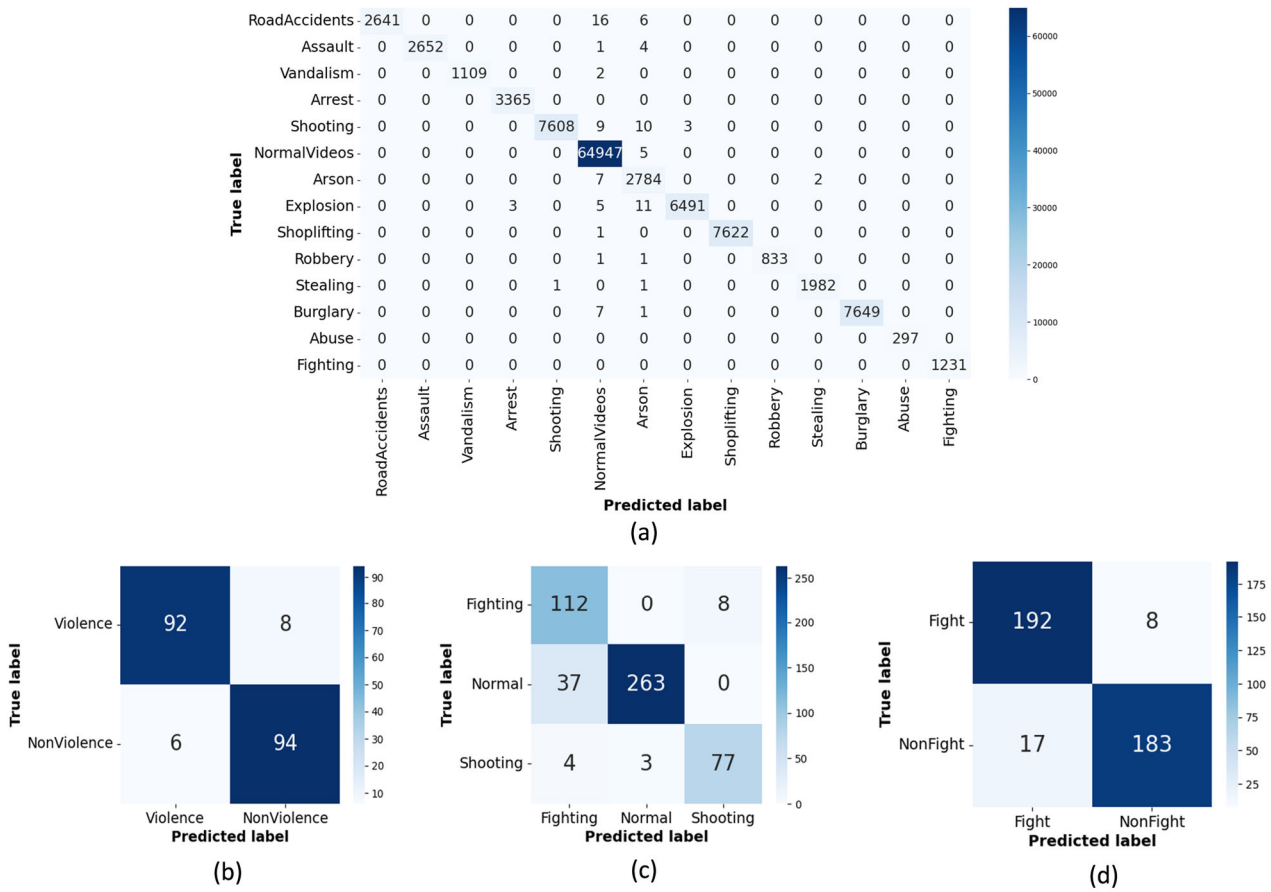

Fig. 6. Confusion matrix of proposed method: (a). UCF-Crime dataset, (b). RLVS dataset, (c). XD_Violence video dataset, (d). RFW-2000 dataset.

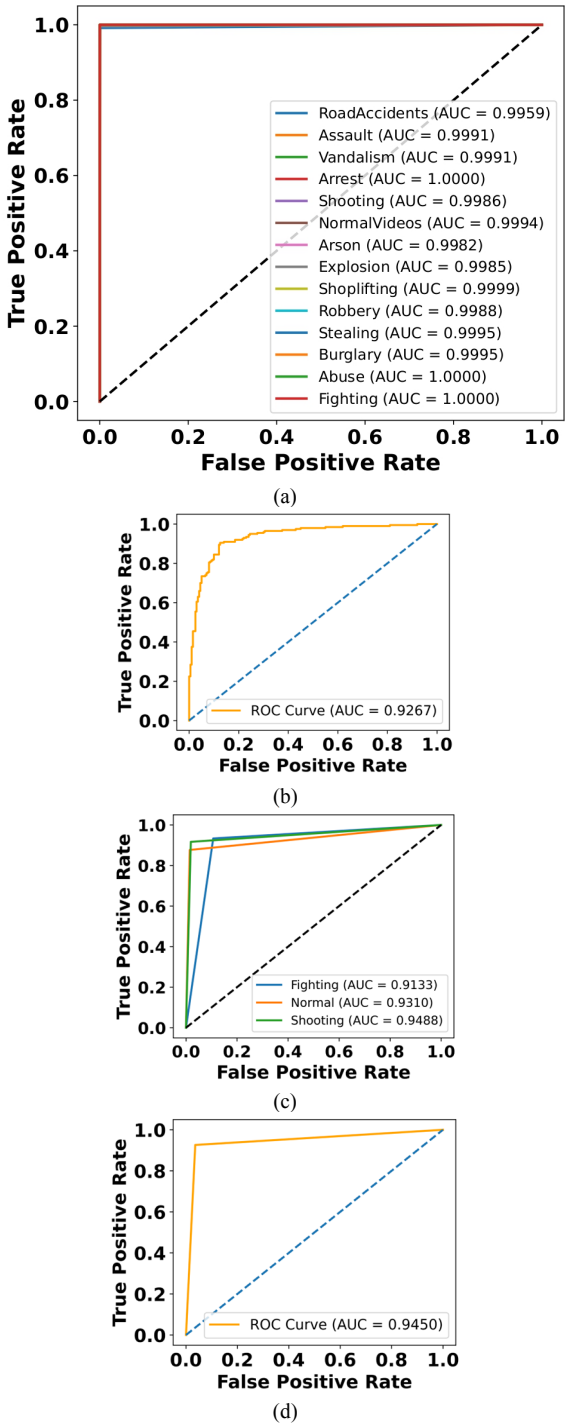

Fig. 7. ROC curve of the proposed method: (a) UCF-Crime dataset, (b) RLVS dataset, (c) XD_Violence video dataset, (d) RFW-2000 dataset.

B. Robustness Analysis Under Realistic Conditions

Although the benchmark datasets utilized in this research are obtained from real-world CCTV and surveillance environments, the robustness experiments were performed by synthetically introducing controlled noise to simulate adverse operational conditions commonly encountered in practical deployments. These conditions include Gaussian noise, salt-and-pepper noise, motion blur, illumination variation, and compression artifacts, which emulate challenges such as sensor noise, camera motion, poor lighting, and video compression in real surveillance systems. The experiments were conducted on UCF-Crime, RLVS, XD-Violence, and RWF-2000 datasets. Table VIII illustrates the robustness evaluation of the proposed CNN-BiLSTM-STAT framework under various realistic noise conditions.

C. Comparative Analysis

This section discusses a comparison between the proposed method and the existing methods. Existing methods such as AVAD [21], SVN [22], deep feature fusion model [23], RedNet50+SRU [24], RNN-LSTM [25], MobileNet [26], hybrid method [27] and TCN [28] were estimated and compared with the proposed approach using various performance metrics based on the four different datasets namely, UCF-Crime, RLVS, XD_Violence video dataset and RFW-2000. Table IX represents a comparative analysis of the proposed QOBL-ARO and GS-SVM.

D. Discussion

Although the proposed CNN-BiLSTM-STAT framework demonstrates superior performance across multiple datasets, but there are several practical and technical considerations that need to be acknowledged. One limitation is the sensitivity of the model to varying video resolutions and lighting conditions, which slightly affects the prediction consistency across environments. Moreover, although the model effectively captures long-term dependencies, the interpretability of deep features remains limited, particularly when distinguishing nuanced anomalous actions from normal crowd behavior. The proposed method significantly outperforms the existing techniques by effectively merging CNN for spatial representation, BiLSTM for bidirectional temporal dependencies, and adaptive attention for dynamic context modeling. This integrated approach enables accurate detection of subtle and complex anomalies and thus improves both precision and recall across diverse video surveillance scenarios. The proposed CNN-BiLSTM-STAT framework illustrates the important enhancements over existing methods by effectively integrating the spatial, temporal, and attention-based modeling for anomaly detection. Beyond achieving a strong benchmark performance, the model also highlights practical relevance through qualitative visualizations, which confirm its ability to localize anomalous regions within video frames. The computational efficiency is important for real-time anomaly detection in continuous video streams.

Although our model reduces training and inference times compared to the other transformer-based methods, further optimization of edge devices and low-power systems will be important for large-scale adoption. Despite its effectiveness, the proposed method exhibits limitations in certain challenging scenarios. Subtle anomalies with minimal motion variation, severe occlusions, and visually ambiguous activities that closely resemble normal behavior occasionally result in misclassification. Additionally, low-resolution or heavily compressed video sequences may degrade feature quality, affecting attention estimation. Addressing these challenges through multi-resolution modeling and context-aware supervision forms

a potential direction for future work. Distributed and cloud-based implementations, combined with parallel processing, can support large-scale monitoring while maintaining responsiveness.

TABLE VIII. ROBUSTNESS EVALUATION OF THE PROPOSED CNN-BILSTM-STAT FRAMEWORK UNDER VARIOUS REALISTIC NOISE CONDITIONS

Dataset	Noise Type	Noise Level	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
UCF-Crime	No Noise (Original CCTV videos)	–	99.91	99.88	99.82	99.85	100
	Gaussian Noise	$\sigma = 0.01$	99.62	99.54	99.41	99.47	99.82
	Gaussian Noise	$\sigma = 0.05$	98.76	98.59	98.41	98.5	99.11
	Gaussian Noise	$\sigma = 0.10$	97.84	97.63	97.29	97.46	98.35
	Salt & Pepper Noise	1%	99.18	99.04	98.91	98.97	99.52
	Salt & Pepper Noise	3%	98.41	98.22	97.96	98.09	98.94
	Motion Blur	Kernel = 5	98.87	98.66	98.52	98.59	99.08
	Motion Blur	Kernel = 9	97.96	97.71	97.38	97.54	98.41
	Illumination Variation	Low-light ($\gamma = 0.7$)	98.69	98.43	98.27	98.35	99.02
	Compression Artifacts	H.264 (High)	98.12	97.88	97.63	97.75	98.56
Real-Life Surveillance Videos (RLVS) Dataset	No Noise (Original CCTV videos)	–	93.00	93.02	93.00	93.00	92.67
	Gaussian Noise	$\sigma = 0.01$	92.64	92.43	92.28	92.35	92.14
	Gaussian Noise	$\sigma = 0.05$	91.84	91.62	91.47	91.54	91.32
	Gaussian Noise	$\sigma = 0.10$	90.71	90.44	90.21	90.32	90.58
	Salt & Pepper Noise	1%	92.12	91.93	91.78	91.85	91.74
	Salt & Pepper Noise	3%	91.06	90.88	90.71	90.79	90.84
	Motion Blur	Kernel = 5	91.96	91.78	91.62	91.70	91.91
	Motion Blur	Kernel = 9	90.32	90.15	89.96	90.05	90.11
	Illumination Variation	Low-light ($\gamma = 0.7$)	91.21	91.03	90.88	90.95	91.02
	Compression Artifacts	H.264 (High)	90.68	90.42	90.21	90.31	90.55
XD-Violence Dataset (Unconstrained Real-world Videos)	No Noise (Original CCTV videos)	–	89.68	87.55	90.89	88.70	93.10
	Gaussian Noise	$\sigma = 0.01$	88.94	86.81	90.11	88.43	92.17
	Gaussian Noise	$\sigma = 0.05$	87.92	85.81	89.14	87.44	91.28
	Gaussian Noise	$\sigma = 0.10$	86.53	84.36	87.59	85.95	90.14
	Salt & Pepper Noise	1%	87.48	85.32	88.71	86.98	90.92
	Salt & Pepper Noise	3%	86.74	84.66	88.02	86.31	90.41
	Motion Blur	Kernel = 5	87.03	84.92	88.34	86.6	90.67
	Motion Blur	Kernel = 9	85.96	83.91	87.37	85.6	89.76
	Illumination Variation	Low-light ($\gamma = 0.7$)	87.38	85.32	88.75	86.99	90.92
	Compression Artifacts	H.264 (High)	86.21	84.08	87.44	85.72	89.93
RWF-2000 Dataset (Real-world Crowd Surveillance)	No Noise (Original CCTV videos)	–	93.75	93.84	93.75	93.75	94.50
	Gaussian Noise	$\sigma = 0.01$	93.21	93.08	92.96	93.02	94.02
	Gaussian Noise	$\sigma = 0.05$	92.61	92.48	92.35	92.41	93.42
	Gaussian Noise	$\sigma = 0.10$	91.68	91.54	91.38	91.46	92.63
	Salt & Pepper Noise	1%	92.94	92.79	92.66	92.72	93.81
	Salt & Pepper Noise	3%	91.89	91.76	91.63	91.69	92.78
	Motion Blur	Kernel = 5	92.31	92.16	92.01	92.08	93.34
	Motion Blur	Kernel = 9	91.12	90.94	90.81	90.87	92.14
	Illumination Variation	Low-light ($\gamma = 0.7$)	92.03	91.88	91.74	91.81	93.01
	Compression Artifacts	H.264 (High)	91.44	91.29	91.12	91.20	92.47

TABLE IX. COMPARATIVE ANALYSIS OF PROPOSED METHOD WITH EXISTING METHODS

Method	Dataset	AUC (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AVAD [21]	UCF-Crime	97.3	-	-	-	-
SVN [22]	UCF-Crime	98.4	-	-	-	-
Deep Feature Fusion Model [23]	UCF-Crime	-	83.59	-	-	-
	RLVS	-	91.99	-	-	-
RedNet50+SRU [24]	UCF-Crime	91.64	91.44	91.71	-	91.64
RNN-LSTM [25]	UCF	-	92.35	94.13	91.44	92.76
Mobile Net [26]	UCF-Crime	-	98.47	-	-	-
Hybrid Method [27]	UCF-Crime	90.09	-	-	-	-
	XD Violence	86.48	-	-	-	-
TCN [28]	UCF-Crime	-	84	-	-	-
	RLVS	-	91.15	-	-	-
	RWF2000	-	93.15	-	-	-
Proposed CNN-BiLSTM-STAT	UCF-Crime	100	99.91	99.88	99.82	99.85
	RLVS	92.67	93.00	93.02	93.00	93.00
	XD_Violence_video	93.10	89.68	87.55	90.89	88.70
	RWF2000	94.50	93.75	93.84	93.75	93.75

E. Failure Case Analysis

The proposed CNN-BiLSTM-STAT approach attains effective performance on standard datasets; however, some failure cases are observed, as shown in the Fig. 8. Misclassifications are mostly present in scenarios with severe occlusion, abrupt camera motion, very low illumination, and visually subtle anomalies that closely resemble normal activities. In crowded scenes, overlapping actions sometimes resulted in ambiguous motion patterns, causing confusion between normal and anomalous events. These observations indicate that performance slightly degrades in highly noisy real-world surveillance conditions, which can be addressed in future work through noise-robust training and improved transformer optimization.

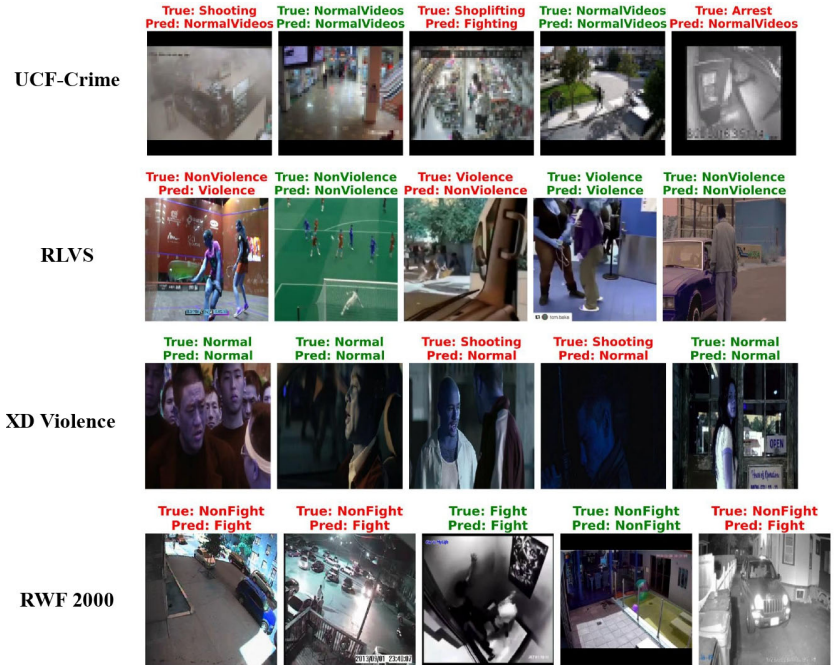

Fig. 8. Failure case analysis.

F. Threats to Validity

In this research, the CNN-BiLSTM-STAT approach is proposed for video anomaly detection. The proposed method attains important findings in anomaly detection as compared with the existing techniques based on the different benchmarks such as UCF-Crime, RLVS, XD-Violence, and RWF2000. However, the findings of the proposed approach may vary because of different factors such as: (1) different pre-processing methods, (2) differences in experimental settings, (3) differences in hyperparameter settings, (4) variances in the base model, and (5) different cross-fold analysis. In future work, this research aims to further evaluate the effectiveness of the proposed CNN-BiLSTM-STAT approach by addressing the previously discussed limitations.

V. Conclusion

Anomalies frequently occur in real-world environments and early detection is critical for ensuring public safety. Identifying such irregular events through surveillance videos plays a key role in ensuring a secure and protected community setting. Although numerous methods have been proposed that use various deep learning and computer vision techniques, many have failed to deliver consistent and reliable results. To address these limitations, this study introduced a CNN-BiLSTM-STAT framework for video anomaly classification in surveillance systems. The proposed model incorporates a generalized spatio-temporal attention mechanism that not only enhances anomaly detection but also serves as a flexible foundation for broader video analysis tasks. The significant contribution of this study lies in illustrating that hybrid deep learning architectures can provide both accuracy and computational efficiency. By minimizing reliance on manual monitoring and enabling timely recognition of unusual events, the proposed method supports safer urban and industrial environments. The results demonstrate that the proposed approach significantly outperforms the existing methods in accurately identifying the anomalous events. The experimental results demonstrate that the proposed CNN-BiLSTM-STAT approach attains an optimal accuracy of 99.91% on the UCF-Crime dataset compared to the existing methods, such as RNN-LSTM and MobileNet. However, only the benchmark dataset was used in this study owing to the lack of access to real-time data. Hence, future work will concentrate on evaluating the proposed framework using real-time data obtained from live surveillance streams to validate its effectiveness in dynamic and real-world settings. Moreover, this research will be extended to incorporate more advanced transformer-based architectures, including large-scale vision transformers and hybrid spatio-temporal transformers, to further enhance long-range dependency modeling.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Divya Uluvaru Hoovayya conceived the research, designed the methodology, and wrote the paper; Josephine Prem Kumar conducted the experiments and analyzed the data; Heena Kousar assisted in implementation, literature review, and manuscript editing; all authors approved the final version.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, East Point College of Engineering and Technology, Bengaluru, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India, for providing the necessary facilities and support to carry out this research work.

REFERENCES

[1] H. Ma and L. Zhang, "Attention-based framework for weakly supervised video anomaly detection," *J. Supercomput.*, vol. 78, no. 6, pp. 8409–8429, 2022.

[2] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Pers. Ubiquitous Comput.*, vol. 28, no. 1, pp. 135–151, 2021.

[3] D. K. Sampath and K. Kumar, "Abnormal crowd behaviour detection in surveillance videos using spatiotemporal inter-fused autoencoder," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 6, pp. 470–481, 2023.

[4] M. Mozaffari, K. Doshi, and Y. Yilmaz, "Self-supervised learning for online anomaly detection in high-dimensional data streams," *Electronics*, vol. 12, no. 9, 1971, 2023.

[5] A. Deshpande, K. Warhade, and P. Sanap, "Abnormal activity recognition with residual attention-based ConvLSTM architecture for video surveillance," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 6, pp. 718–730, 2023.

[6] Y. Wang, T. Liu, J. Zhou, and J. Guan, "Video anomaly detection based on spatio-temporal relationships among objects," *Neurocomputing*, vol. 532, pp. 141–151, 2023.

[7] W. Liang, J. Zhang, and Y. Zhan, "Weakly supervised video anomaly detection based on spatial-temporal feature fusion enhancement," *Signal, Image Video Process.*, vol. 18, no. 2, pp. 1111–1118, 2023.

[8] M. Shoaib, A. Ullah, I. A. Abbasi, F. Algarni, and A. S. Khan, "Augmenting the robustness and efficiency of violence detection systems for surveillance and non-surveillance scenarios," *IEEE Access*, vol. 11, pp. 123295–123313, 2023.

[9] S. Ul Amin, Y. Kim, I. Sami, S. Park, and S. Seo, "An efficient attention-based strategy for anomaly detection in surveillance video," *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, pp. 3939–3958, 2023.

[10] A. Zahra, M. Ghafoor, K. Munir, A. Ullah, and Z. Ul Abideen, "Application of region-based video surveillance in smart cities using deep learning," *Multimedia Tools Appl.*, vol. 83, no. 5, pp. 15313–15338, 2021.

[11] A. Elmetwally, R. Eldeeb, and S. Elmougy, "Deep learning based anomaly detection in real-time video," *Multimed. Tools Appl.*, vol. 84, no. 11, pp. 9555–9571, 2025.

[12] I. A. Alnajjar, L. Almazaydeh, A. A. Odeh, A. A. Salameh, K. Alqarni, and A. A. B. Atta, "Anomaly detection based on hierarchical federated learning with edge-enabled object detection for surveillance systems in industry 4.0 scenario," *International Journal of Intelligent Engineering & Systems*, vol. 17, no. 4, pp. 649–665, 2024.

[13] U. Arul, V. Arun, T. P. Rao, R. Baskaran, S. Kirubakaran, and M. T. Hussan, "Effective anomaly identification in surveillance videos based on adaptive recurrent neural network," *Journal of Electrical Engineering & Technology*, vol. 19, no. 3, pp. 1793–1805, 2024.

[14] J. Amin, M. A. Anjum, K. Ibrar, M. Sharif, S. Kadry, and R. G. Crespo, "Detection of anomaly in surveillance videos using quantum convolutional neural networks," *Image Vision Comput.*, vol. 135, 104710, 2023.

[15] J. Chen, B. Liu, and H. Zuo, "Abnormal behavior detection in industrial control systems based on CNN," *Alexandria Eng. J.*, vol. 107, pp. 643–651, 2024.

[16] M. Nallappan and R. Velswamy, "Exploring deep learning-based content-based video retrieval with hierarchical navigable small world index and ResNet-50 features for anomaly detection," *Expert Syst. Appl.*, vol. 247, 123197, 2024.

[17] M. Yan, Y. Xiong, and J. She, "Memory clustering autoencoder method for human action anomaly detection on surveillance camera video," *IEEE Sens. J.*, vol. 23, no. 18, pp. 20715–20728, 2023.

[18] A. Mumtaz, A. B. Sargano, and Z. Habib, "Robust learning for real-world anomalies in surveillance videos," *Multimedia Tools Appl.*, vol. 82, no. 13, pp. 20303–20322, 2023.

[19] J. Arunnehru, "Deep learning-based real-world object detection and improved anomaly detection for surveillance videos," *Mater. Today Proc.*, vol. 80, pp. 2911–2916, 2023.

[20] S. Vosta and K. C. Yow, "KianNet: A violence detection model using an attention-based CNN-LSTM structure," *IEEE Access*, vol. 12, pp. 2198–2209, 2023.

[21] M. M. Ali, "Real-time video anomaly detection for smart surveillance," *IET Image Proc.*, vol. 17, no. 5, pp. 1375–1388, 2022.

[22] Y. Liu, D. Yang, G. Fang, Y. Wang, D. Wei, M. Zhao, K. Cheng, J. Liu, and L. Song, "Stochastic video normality network for abnormal event detection in surveillance videos," *Knowledge-Based Syst.*, vol. 280, 110986, 2023.

[23] S. A. Jebur, L. Alzubaidi, A. Saihood, K. A. Hussein, H. K. Hoomod, and Y. Gu, "A scalable and generalised deep learning framework for anomaly detection in surveillance videos," *Int. J. Intell. Syst.*, vol. 2025, no. 1, 1947582, 2025.

[24] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," *Results Eng.*, vol. 18, 101026, 2023.

[25] V. A. Kotkar and V. Sucharita, "Fast anomaly detection in video surveillance system using robust spatiotemporal and deep learning methods," *Multimedia Tools Appl.*, vol. 82, no. 22, pp. 34259–34286, 2023.

[26] M. Rani and M. Kumar, "MobileNet for human activity recognition in smart surveillance using transfer learning," *Neural Comput. Appl.*, vol. 37, no. 5, pp. 3907–3924, 2024.

[27] J. Shin, Y. Kaneko, A. S. M. Miah, N. Hassan, and S. Nishimura, "Anomaly detection in weakly supervised videos using multistage graphs and general deep learning based spatial-temporal feature enhancement," *IEEE Access*, vol. 12, pp. 65213–65227, 2024.

[28] W. Ullah, F. U. M. Ullah, Z. A. Khan, and S. W. Baik, "Sequential attention mechanism for weakly supervised video anomaly detection," *Expert Syst. Appl.*, vol. 230, 120599, 2023.

[29] UCF-Crime dataset. (2025). [Online]. Available: https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset

[30] RLVS dataset. (2025). [Online]. Available: https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset/data.

[31] XD_Violence video dataset. (2025). [Online]. Available: https://www.kaggle.com/datasets/nguhaduong/xd-violence-video-dataset.

[32] RWF2000 dataset. (2025). [Online]. Available: https://www.kaggle.com/datasets/vulamnguyen/rwf2000.

[33] S. Saifullah and R. Dreżewski, "Modified histogram equalization for improved CNN medical image segmentation," *Procedia Comput. Sci.*, vol. 225, pp. 3021–3030, 2023.

[34] W. A. H. Aljuaid and S. S. Alshamrani, "A deep learning approach for intrusion detection systems in cloud computing environments," *Applied Sciences*, vol. 14, no. 13, 5381, 2024.