Event based Video Summarization for Traffic Surveillance

A PhD synopsis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer/IT Engineering

By

Chauhan Ankita P.

[199999913508]

under Supervision of

Dr. Sudhir P Vegad,
Professor and Head,
Department of Computer Engineering
G H Patel College of Engineering and Technology, Vallabh Vidyanagar University

GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD

[April - 2026]

TABLE OF CONTENTS

List of Figures

List of Tables

1. Title of the Thesis and Abstract

Title of the Thesis:

Event based Video Summarization for Traffic Surveillance

Abstract:

Due to the high rate of development of traffic surveillance systems, continuous video streams are becoming harder to monitor manually. The task of detecting rare accident events and creating compact video summaries efficiently still is a difficult one. The paper suggests a single framework to detect unsupervised accidents and video summarization with deep autoencoders-based structures.

The framework proposed has two models: a Convolutional Variational Autoencoder-based Accident Detection and Summarization (CVAE-ADS) model and a Convolutional Bidirectional Long Short-Term Memory Autoencoder (Conv-BiLSTM AE) model. Both methods will be able to learn spatio-temporal representations of video data without using labelled annotations. CVAE-ADS model identifies the anomalies through the behavior of reconstructions and captures the latent distributions of normal patterns. Conv-BiLSTM autoencoder goes even a step further to improve the temporal feature learning, through the modelling of the bi-directional dependencies between the frame sequences. The latent space clustering mechanism is used to choose representative frames, which makes video summarization possible.

The suggested models are tested on IITH and UCF-Crime data sets. The experimental outcomes show better detection performance than baseline models with high accuracy (up to 94.0%), better AUC (up to 0.946) and lower Equal Error Rate (EER 14.97%). Moreover, the framework attains effective video summarization up to a reduction of 85 percent whilst retaining excellent visual quality (PSNR greater than 30 dB) and variety in frames chosen. In general, the suggested methods are a powerful and effective solution to accident detection and video summarization, with a good trade-off between their detection capability, compression performance and reconstruction quality, which can be applied to the real-life applications of intelligent surveillance.

Keywords: Accident Detection, Video Summarization, Convolutional Variational Autoencoder (CVAE), Conv-BiLSTM Autoencoder, Latent Space Clustering, Intelligent Surveillance Systems

2. Brief Introduction

The past few years have witnessed the rapid growth of urban infrastructure and increase in the number of vehicles, which has greatly complicated the traffic management systems. With the development of cities into smart environments, there has never been a greater demand on efficient and intelligent surveillance of traffic. The surveillance cameras are extensively used in the urban road systems and large volumes of video data are generated which hold useful information regarding traffic movement, incidents and irregularities. Nevertheless, due to manually analyzing such large data, it becomes very not only time consuming but can lead to errors made by humans, thus there is a need to develop automated and intelligent systems [1].

Technological progress in AI, communication technologies and edge computing have made it possible to create video analytics systems for real world traffic syatem that enhance scalability and responsiveness in the current surveillance environment [2]. Moreover, the inclusion of IoT-based models and enhanced object detection models has helped to monitor traffic automatically and respond to emergencies [3]. Moreover, connected vehicle data have also led to better traffic analysis using the detection of near-crash events and further protection of road safety [4]. Predictive models based on DL have as well shown a lot of potential in understanding and predicting complicated traffic trends and enhancing the accident detection and prevention system [5]. Although these developments are made, there are still a number of issues such as proper identification of rare accident occurrences, efficient handling of massive video data, and the creation of succinct summaries that will point out major accidents.

In addition, the majority of current methods are concentrated on the individual tasks, i.e., accident detection or video summarization in isolation, but not on the combination of such tasks. It leaves a vacuum in its ability to come up with effective systems that can not only monitor incidents but also give meaningful summaries to enable prompt analysis and action. Thus, it is highly desirable to have powerful structures capable of integrating detection and summarization and being scalable, accurate, and real-time. In this regard, the current study will seek to create a state-of-the-art DL-based traffic surveillance system that will be used to detect traffic accidents and video summarization events. The suggested solution employs the unsupervised learning methods and hybrid DL models to overcome the constraints of the current systems and make a step towards the creation of intelligent and efficient traffic monitoring systems.

2.1 Research Focus

As cities continue to expand and increase in population and size, traffic monitoring systems are becoming an essential part of a smart city. The surveillance cameras create enormous video data that is difficult to analyze manually and inefficient. As a result, intelligent video analytics has become an important area of research to derive meaningful information like the ability to detect accidents, anomalies and summarize events in the large-scale video streams [1].

Recently developed edge computing and 5G-enabled systems have further increased the possibility of processing traffic video data on-the-fly, enhancing scalability and responsiveness of surveillance systems [2]. Moreover, IoT-based systems with object recognition models have facilitated automated traffic surveillance and response systems in case of an emergency [3] [4].

Predictive models that use deep learning have shown detection with high accuracy and prevention of accidents through acquisition of intricate patterns using traffic data [5]. Moreover, hybrid techniques using both geographic and social data sources have been considered to enhance the prediction of the severity of accidents and the decision-making process [6]. There have also been vision based systems that are developed to detect unsafe driving habits and violations of the traffic code as well as helping in improving safety on the road [7].

A thorough review of traffic monitoring technologies reveals that it is necessary to combine detecting and summarization methods to have effective systems of monitoring [8]. The present advanced deep learning systems, such as those based on the YOLO, have also enhanced real-time performance of accident detection [9]. Also, smart traffic systems that incorporate incident detection and vehicle alert systems are rapidly becoming part of smart city landscape [10].

Recent studies also pay attention to the early accident risk prediction based on cross-modal learning and collaborative perception framework as they increase the predictive accuracy [11]. In addition, there are new video storage and processing technologies that enable the efficient processing of continuous streams of surveillance data in large scale [12].

To gain a clearer idea of how traffic surveillance systems have developed, and to determine the shortcomings of current methods, Table 1 presents a comparative analysis of the various methods and their features.

Based on the discussion and analysis above, it is clear that in spite of the immense progress made, most of the current approaches tend to concentrate on either accident detection or video

Table 1 Development of Traffic Surveillance and Accident Detection methods

Approach Type	Key Technique Used	Strengths	Limitations
Traditional Methods	Manual monitoring, rule-based systems	Simple implementation	Time-consuming and error-prone
Machine Learning-Based [6]	SVM, Random Forest, feature-based models	Moderate accuracy	Requires feature engineering
Deep Learning-Based [5][9]	CNN, YOLO models	High detection accuracy	Limited temporal understanding
Spatiotemporal Models [11]	CNN + LSTM / BiLSTM	Captures temporal dependencies	High computational cost
IoT & Edge-based Systems [1][2][3]	Sensors + edge computing	Real-time processing	Limited summarization capability
Large-scale Video Systems [12]	Cloud-based video storage	Handles big data efficiently	Storage overhead
Proposed Research Direction	CVAE-ADS and Conv-BiLSTM Autoencoder based	Integrated detection & summarization	Needs real-time optimization

summarization on a case-by-case basis. Thus, the proposed research paper aims at creating a unified event-based video summarization framework of traffic surveillance that will integrate the accident detection and event summarization into a single system. The suggested work employs unsupervised deep learning models, including Convolutional Variational Autoencoders (CVAE) and Conv-BiLSTM Autoencoders, to find anomalies and create short descriptions of crucial events, thus enhancing the overall surveillance efficiency.

2.2 Problem Statement

The high rate at which the number of vehicles is increasing has largely been a contributor to the increase in road accidents, which poses a great challenge to the traffic monitoring and management system. Despite the high usage of surveillance cameras, it is a challenging and resource-consuming task to extract meaningful information using the continuous video streams. The correct identification of the accident events is one of the greatest challenges because they are infrequent, unforeseeable and usually happen in complicated environment conditions, like occlusion, low visibility and dynamic traffic situations [13]. Current systems may fail to detect the exact time of an accident which leads to low detection performance.

The other major challenge is the fact that the existing systems have little ability to identify abnormalities in different traffic conditions [14]. Despite the enhanced accuracy of vision-based classification models in detecting objects, they frequently are unable to provide the contextual and behavioral patterns that are needed to detect accidents with reliable accuracy [15]. Besides, shadows, changes in illumination, and noise are environmental factors that greatly affect the performance of deep learning models [16]. Detection of anomaly in surveillance systems has been difficult because of the variety and randomness of abnormal events [17].

Moreover, the severity and effects of accidents are not simple to understand and can only be done using high-level predictive modeling [18]. Even the current methods of image detection have limitations in terms of detecting collision events in real-time situations [19]. In addition, real time detection of accidents with high reliability and efficiency is still a significant challenge to the existing systems [20]. To systematically point out these problems, Table 2 summarizes the main challenges in traffic surveillance systems and respective research requirements.

Table 2: Challenges and Research Needs in Traffic Surveillance

Challenge	**Description**	**Research Need**
Large Video Data [8][12]	Continuous surveillance generates massive data	Automated scalable analytics
Rare Accident Events [20]	Accidents are infrequent and unpredictable	Robust anomaly detection
High False Alarms [17]	Misclassification of normal events	Improved detection accuracy
Lack of Temporal Modeling [5][11]	Frame-level analysis only	Sequence-based models
Redundant Video Content [14][15]	Repetitive frames reduce efficiency	Effective summarization
Complex Conditions [13][16]	Occlusion, low-light issues	Robust detection models
Severity Analysis [18][19]	Difficulty in impact assessment	Advanced prediction models
Labelled Data Dependency [6]	Requires manual annotation	Unsupervised learning

According to the challenges presented above, it is highly desirable to have a common framework that can effectively identify accident events, record spatiotemporal features, minimize false alarms, and produce succinct summaries of significant events effectively. This study discusses these problems and proposes a unified deep learning-based accident detection and event-based video summary system.

2.3 Motivation

The rationale behind this study is that there has been an escalating need to adopt automated and intelligent traffic monitoring systems that can contribute to road safety and efficiency in traffic control. As CCTV cameras and smart city infrastructure are deployed at a rapid pace it is vital to be able to convert raw video data into usable insights as quickly as possible. The need to go beyond the constraints of the old systems that are largely manual-based and labelled datasets is one of the overriding reasons. The unsupervised learning techniques offer a viable solution as they allow models to learn normal traffic behaviour, as well as identify anomalies, without a lot of annotation.

The other motivation is to close the divide between the detention of accidents and video summarization that are frequently viewed as distinct processes in the literature. The combination of these features within one structure greatly enhances the effectiveness and ease of use of the systems. Moreover, the development of deep learning architectures, especially hybrid networks that are convolutional and recurrent networks, permit the successful modelling of both spatial and temporal dependencies of video data. This inspires the discussion of CVAE and Conv-BiLSTM models to enhance the level of detection and the quality of summarization.

The need to create computationally efficient systems capable of functioning in real-time and edge computing architectures also drives the research. Surveillance data processing should be efficient to enable prompt decision making in traffic monitoring and emergency response systems. The ultimate goal of this work is to make contributions that will lead to the creation of intelligent, scalable and real-time traffic monitoring systems that save human resources, enhance accuracy in detecting accidents, offer meaningful summaries and assist authorities in decision-making, which will result in a better road safety and flow in the city.

3. Literature Survey and Research Gap

The surge of demand on road safety and effective traffic control has resulted in serious research in recent years which aimed at creating intelligent traffic surveillance systems. As the field of artificial intelligence, computer vision, and deep learning methods rapidly evolves, a variety of methods have been suggested to solve the problems of traffic accident detection, anomaly recognition, and video analysis. These solutions are designed to provide automation of monitoring of vast amounts of surveillance data and help in timely decision-making in traffic authorities. Nevertheless, the range of different methodologies and the reality of real-life traffic situations have resulted in different degrees of performance and applicability in different systems [21][23].

Additionally, the inclusion of cutting-edge solutions like multimodal data analysis, IoT-based solutions, and real-time video analytics has also contributed to the further development of traffic surveillance systems. A number of studies have aimed to enhance the level of accuracy in detecting objects via deep learning models, whereas others have also sought to implement tracking-based, data-driven, and hybrid methods to understand traffic behavior [25][30][31]. Even with these developments, most of the current techniques are still challenged by issues of scalability, computational performance, and capability to address intricate spatiotemporal trends in traffic videos.

In that respect, it is necessary to examine the available literature to see the current level of research, to see what limitations the current methodologies have, and to lay the groundwork to the proposed work. Consequently, the section below introduces a thorough discussion of the existing research works in the field of traffic monitoring, crash detection, and video summarization, as well as the identification of potential gaps in research.

3.1 Literature Survey

The latest developments in smart traffic monitoring systems have greatly improved the ability to monitor, analyze and react to road accidents. The methodologies examined in different studies include edge computing and multimodal analytics to deep learning-based accident detection and tracking systems. Multimodal frameworks, edge-enabled traffic video analytics have facilitated real-time processing and enhanced scalability of surveillance systems [1][2]. Emergency response systems built on IoT also enhance the control over traffic, allowing automated incident detection and response [3]. Moreover, near-crash detection and accident prevention using connected vehicle data and predictive models have been applied in the urban setting [4][5]. Hybrid methods using both geographic and contextual data have also been discussed to enhance the prediction of the severity of accidents [6].

Surveillance systems based on vision have extensively been used in tracking traffic operations such as unsafe road use and detection of violations [7]. The use of detection and summarization methods in the development of efficient traffic monitoring is emphasized in extensive surveys [8]. CNN and YOLO-based architectures in deep learning models have been shown to perform well in real time accident detection tasks [9][10]. Moreover, to improve early accident risk prediction, cross-modal perception and collaborative learning systems have been proposed [11], and large-scale video processing systems have been implemented to provide efficient storage and analysis of surveillance data [12].

Some works have been devoted to the anomaly detection and classification in surveillance videos in adverse environments like occlusions and dynamic scenes [13][14][15]. The optimization of advanced features and adaptive learning methodologies have been suggested to enhance the performance of anomaly recognition [16][17]. Also, predictive modeling methods have been discussed to analyze the severity of accidents and enhance their detection accuracy [18][19]. Real time detection systems have also been invented to improve road safety and emergency response [20].

Most recent literature also discusses more sophisticated object tracking and generative models to use on surveillance purposes. Speed detection algorithms used in automated traffic control systems can help to improve proactive management of traffic through motion analysis and tracking accuracy, which is optimized using GAN-based tracking methods [21]. Correlation-based tracking methods have proven to be effective in detecting accidents in video streams [23], and comparative analysis of machine learning methods reveals the necessity to choose an effective tracking model [24].

Comprehensive traffic incident response systems have been suggested using multimodal systems that combine video, sensor and contextual data [25]. There have also been lightweight deep learning models to detect anomalies in various modalities, including sound-based highway monitoring [26]. CNN architecture-based deep learning-based accident detection systems have demonstrated better accident detection performance in surveillance settings [27], and fuzzy logic-based systems have been investigated in the context of vehicular networks [28]. Intelligent traffic lights that have inbuilt anomaly detecting systems also help in effective management of traffic [29].

Ensemble learning and data-driven methods have increased the generalization performance and robustness of accident detection systems [30][31]. Development of real-time CCTV-based accident detection and emergency response systems has also been done [32]. People and vehicle tracking systems based on AI are common in surveillance applications [33], whereas trajectory-based and influence mapping techniques enhance the detection accuracy of accidents [34]. Vehicle detection methods powered by machine learning enhance traffic monitoring systems to a greater degree [35].

Techniques based on feature extraction which involve the characteristics of traffic flow have been suggested to enhance the detection of accidents [36], and the computer vision systems have shown to be promising in the classification of traffic incidents in real-time [37]. There is an introduction of synthetic data generation methods to deal with the shortcomings of labelled datasets [38]. CCTV analytics-based event-driven monitoring systems allow real-time traffic analysis [39], whereas edge computing-based systems enhance better processing in distributed environments [40]. Temporal pattern mining techniques have been used to detect the accident prone traffic conditions as well [41].

Table 3 is a comparative summary of the previous studies, their methodologies, experimental procedures, data sets and gaps in the research to systematically analyze the research work and indicate the limitations of the research studies.

Table 3: Summary of Prior Studies and Research Gap

Ref.	**Methodology**	**Experimental Techniques**	**Datasets**	**Research Gap**
[10][11] [12]	Intelligent traffic systems with alerting mechanisms	Cross-modal learning and large-scale video processing	Smart city surveillance datasets	High computational complexity and storage overhead

[13][14] [15]	Video anomaly detection and classification models	CNN + feature extraction + classification	Surveillance video datasets	Ineffective keyframe extraction for summarization
[16][17] [18]	Deep learning-based anomaly recognition and prediction	Feature optimization and adaptive models	Traffic anomaly datasets	High false alarm rates and poor generalization
[19][20] [21]	Accident detection using reconstruction and GAN-based tracking	Autoencoder, GAN, reconstruction error	CCTV accident datasets	Limited temporal dependency modeling
[22][23] [24]	Traffic control and tracking-based detection approaches	Speed detection, correlation tracking	Traffic surveillance datasets	Focus on detection only, no summarization
[25][26] [27]	Multimodal and deep learning frameworks for incident detection	CNN, multimodal fusion, attention mechanisms	Real-world traffic datasets	Lack of unified framework for detection and summarization
[28][29] [30]	Fuzzy logic and data-driven anomaly detection models	Rule-based + ML hybrid techniques	Vehicular network datasets	Low scalability and accuracy issues
[31][32] [33]	Ensemble learning and real-time surveillance systems	Deep ensemble models, CCTV analytics	Real-time surveillance data	Computational complexity and redundancy in output
[34][35] [36][37] [38][39] [40][41]	Trajectory-based, edge computing, and pattern mining approaches	Tracking, temporal mining, CV-based detection	Traffic flow and video datasets	Limited integration of spatiotemporal learning with summarization

3.2 Research Gap

According to the extensive literature review provided in Section 3.1, one can conclude that much has been done in the area of traffic monitoring and accident recognition. Nonetheless, there are a number of severe constraints, which continue to limit the functionality and usability of current systems in practice. The existing methods have mostly been based on the supervised learning algorithms, which need huge amounts of labelled data to be trained. This scalability is constrained particularly in practical surveillance systems where labelled data are rare and costly to create [30][31]. Moreover, most of the current solutions mainly address the individual functions like detection, tracking or prediction functions, without combining these functions into a single framework. Consequently, the potential of meaningful summarization of identified events is not well investigated [39].

The other significant limitation is the poor modelling of time dependencies. Despite the fact that some of them involve sequence-based learning, many methods continue to use frame-based analysis that is unable to capture the dynamics of traffic events [34] [41]. Also, the available systems are usually characterized by a high level of computational complexity, and thus cannot be deployed in real-time, specifically in edge computing environments [40]. Detection systems are also less reliable due to environmental issues like occlusion, illumination differences, and noise [37]. Besides, the problem of redundancy of video data is not effectively solved, which causes inefficient summarization and higher storage needs.

Overall, the review of the literature shows that there are some important research gaps that should be filled.

- The integration between video summarization and accident detection is also greatly lacking, with the majority of the current systems addressing the two processes separately and not as a single process. Also, existing methods are highly reliant on labelled data because of widespread application of supervised learning methods, impairing their scalability and real-world viability.

- Moreover, most models have a poor ability to represent spatial and temporal characteristics which translates to poor spatiotemporal modelling of complex traffic situations. The other significant issue is that the current methodologies are characterized by the high computational complexity of their algorithms and are therefore limited to real-time traffic monitoring systems.

- The redundant video data problem is also not effectively managed, thus creating redundant storage and processing overheads. Furthermore, they lack well-developed unsupervised anomaly detection models that can be effectively used without having to resort to large, annotated datasets.

All these constraints underscore the need to establish a cohesive, effective and scalable intelligent traffic surveillance system that can help resolve these issues comprehensively.

3.3 Research Objective

The main goal of the study is to create an effective and integrated deep learning-based framework of detecting traffic accidents and summarizing videos according to the events. The purpose of the proposed framework is to resolve the shortcomings in the current systems by combining anomaly detection and summarization into one pipeline. The study aims at taking advantage of unsupervised learning algorithms to remove the reliance on labelled data and achieve high detection rates. It is also interested in adding both spatial and temporal feature learning to enhance the comprehension of complex traffic situations.

- There are a number of objectives that have been established to fulfil the overall objective of this research. The former is to develop an effective framework of unsupervised anomaly detection based on Convolutional Variational Autoencoders (CVAE) that can acquire normal traffic distribution and detect anomalous situations without using any labelled data.

- The second purpose is to create a model based on Conv-BiLSTM that can be able to capture the temporal relations of video sequences and can enhance the detection of dynamic traffic events. Moreover, the study will also conduct a comparative analysis of the non-temporal (CVAE-based) and temporal (Conv-BiLSTM-based) methods to analyze their efficiency in accident detection.

- Another critical goal is to introduce an event-based video summarization system based on latent space clustering methods, which allows extracting meaningful keyframes that reflect key events. In addition, the paper aims at minimizing redundancy in surveillance videos by the use of effective keyframe selection and similarity-based filtering mechanisms, which result in the creation of small but informative summaries.

The study will also strive to determine the effectiveness of the suggested models by assessing them using the conventional performance measurement tools like accuracy, precision, recall, F1-score, AUC and EER, which will provide a holistic evaluation of the system. Lastly, another significant aim is to make the design such that it is computationally efficient and can be applicable in real-time, which means it can be deployed in real-life traffic surveillance systems.

3.4 Scope of the Work

This research is limited to creating a scalable and efficient structure of traffic surveillance systems based on video-based data. The paper highlights the application of unsupervised DL methods in the detection of anomalies and brings the accident detection as well as video summarization together as a Unified approach. The IITH Dataset and UCF-Crime Accident subset Dataset serve as benchmark datasets to evaluate the proposed framework taking into account both spatial and temporal modelling. Applicability to real-time and efficiency in computation are also taken into account in realistic limitations of the research.

The focus of the current study is the creation of an effective and scalable architecture for a video based traffic surveillance. The research mainly dwells upon monitoring traffic by CCTV-based surveillance, which is the main data source to analyse and assess.

- The study focuses on application of unsupervised learning methods to detect anomalies, with an aim of minimizing the reliance on labelled data and still perform well. Moreover, the work combines the accident detection and video summarization tasks in one framework and makes sure that detection and summarization are carried out in a single structure, and not separately.
- The proposed models are tested with standard benchmark datasets, including the IIT Hyderabad Dataset and the UCF-Crime Dataset, to make sure that they are reliable and comparable to existing solutions. Besides, the study takes into account both the spatial and the temporal modelling methods to be able to capture the dynamic nature of traffic event.
- Lastly, the study also gives much importance to attaining computational efficiency and scalability, whereby the suggested framework can be implemented in real-time traffic monitoring areas and is capable of processing large volumes of video data.

3.5 Innovative Contributions

The suggested study makes a number of new contributions to the sphere of intelligent traffic surveillance systems. The main novelty of this work is the formation of an integrated structure that will combine the accident detection and video summarization, which are traditionally considered different tasks in the current literature. The proposed solution would improve the effectiveness and practical usefulness of the current traffic surveillance systems by integrating all these capabilities into one pipeline. The other important impact of this study is the use of unsupervised deep learning models, specifically the Convolutional Variational Autoencoder (CVAE) which allows an effective detection of anomalies without the need to use mass labelled data.

Such a strategy enhances scalability and flexibility of the system in real-life situations. Moreover, the model is designed with a Conv-BiLSTM Autoencoder that enables the model to learn temporal relationships among video sequences and thus improve the recognition of dynamic and intricate traffic scenes. Moreover, the study suggests an effective video summarization algorithm that utilizes latent space clustering to extract important keyframes that represent key events. The technique enhances the quality of summaries generated, as well as minimising redundancy by the use of similarity-based filtering methods, which ensure concise and informative summaries.

The research also has its contribution as it provides a comparative analysis of the various approaches to deep learning, especially between non-temporal and temporal models, and gives a priceless idea about how they perform and their applicability to traffic surveillance systems. Generally, the proposed study makes important contributions towards the creation of real-time intelligent, and scalable traffic surveillance systems that can be effective in facilitating better decision-making, road safety, and effective management of traffic in smart city settings.

4. Methodology of Proposed Research

This study suggests an event-based video summarization system of traffic surveillance based on the integration of anomaly detection with DL based models. The methodology is aimed at detecting the accident events in continuous video streams based on reconstruction-based learning. There are two models: the CVAE-ADS and Conv-BiLSTM Autoencoder, which are developed to learn the spatial and spatio-temporal features of the traffic scenes. Anomalies in the system are detected by the reconstruction error and summaries are created by only keeping frames of events that are significant, thus eliminating repetition and still retaining important information.

4.1 Overview of Proposed System

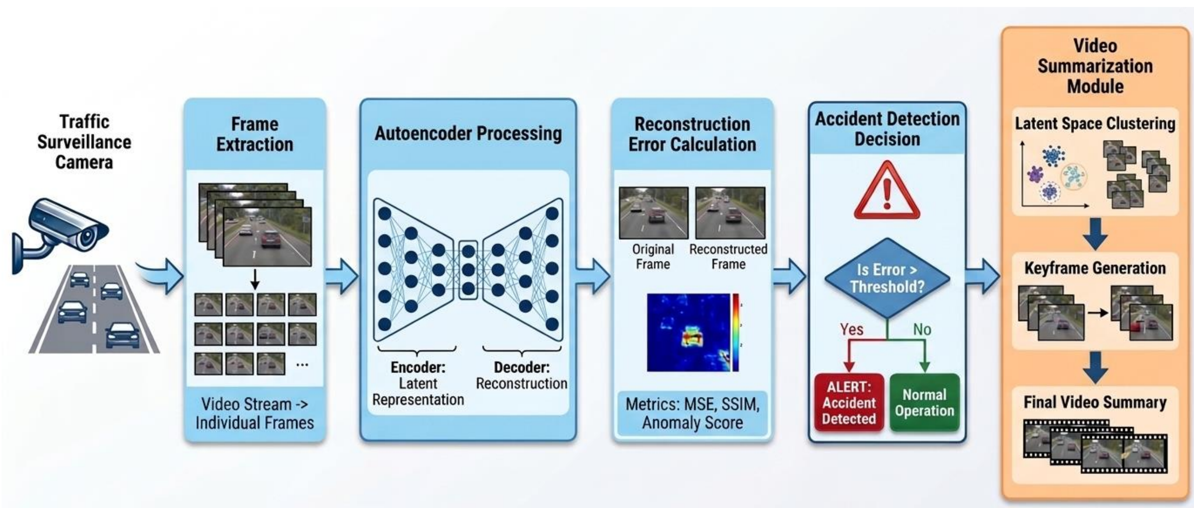


Fig. 1 Overview of the proposed Autoencoder-based traffic accident model.

The proposed system will automatically identify the incidences of accidents and produce summaries of the video records of constant traffic monitoring. The process starts as shown in Fig. 1 by converting the input video stream to individual frames which can be analyzed in detail. These frames are fed through an autoencoder-based model, either as the CVAE-ADS model or the Conv-BiLSTM autoencoder. The model is trained on the regular traffic dynamics and recreates the input frames. In the occurrence of abnormal events like accidents, the reconstruction quality is compromised leading to increased reconstruction error.

The reconstruction error between the original and reconstructed frames is calculated to detect such anomalies, which is based on a chosen metric. In this work, Mean Squared Error (MSE) is used for efficient error estimation. A decision mechanism is a threshold-based mechanism that classifies the frame as either normal or abnormal with frames above the threshold being classified as accident-related.

To summarize, the detected abnormal frames are then further summarized in a video summarization module. Clustering is then used to group the corresponding latent representations into similar event patterns and keyframe extraction is then used to choose representative frames. These keyframes are subsequently placed in sequence to produce the end result of the summarized video, in such a way as to achieve a concise but informative account of the accidents.

4.2 Dataset Description and Preprocessing

Experiments are performed on two publicly available datasets: the IIT Hyderabad Dataset [42] and the UCF-Crime (Accident subset) [43] to test the proposed framework. These datasets present various real-life traffic conditions and are popular in carrying out anomaly detection tasks in surveillance videos.

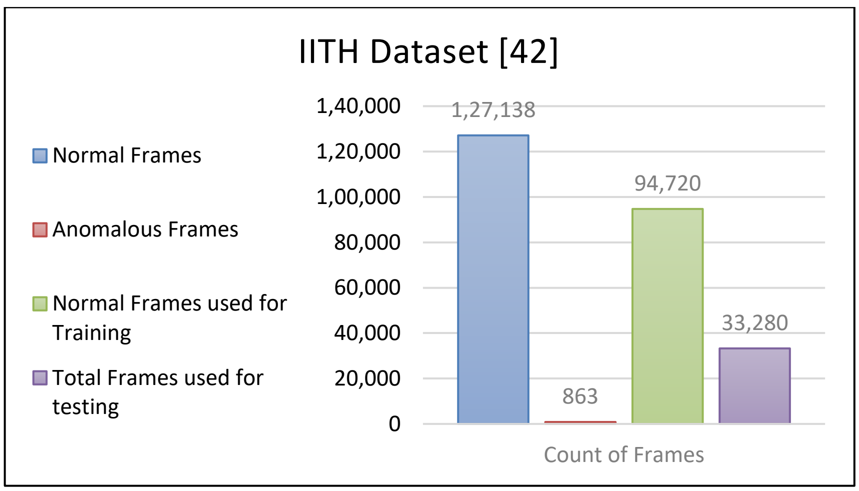


Fig. 2 Details of IITH Road Accident Dataset

The IITH Dataset [42] is a collection of videos that were recorded on CCTV surveillance cameras that were installed in Hyderabad city, India. The videos are captured at 30 frames per second and normally start with the normal traffic conditions and then the incidence of an accident event. In this work, the first part of every video, where there is only normal traffic behaviour, is utilized to train the model to learn common behaviour. The other section, which contains the accident events, is utilized to test. The statistics of normal and anomalous samples at the frame level applied in training and testing is summarized in Fig. 2. Fig. 3 demonstrates the representative frames of the IITH road accident dataset [10], depicting the common traffic conditions and accident situations, which are used in training and testing.

The UCF-Crime dataset [43] is a large scale benchmark dataset that is created to detect anomalies

Fig. 3 Video frame samples of IITH road accident data set [42].

in surveillance videos. It has about 1,900 long raw videos of about 128 hours of content that includes 13 types of real-world anomalies including road accidents, robbery, fighting and vandalism. In this piece, the subset of accidents relevant to this work, 151 videos, is used. These videos are difficult real-life scenarios such as occlusions, light changes, and different camera angles. This dataset does not have pre-defined frame-level training and testing splits as IITH does. Thus, a uniform preprocessing policy is used, with frames being picked and categorized according to time blocks to provide equal consideration.

4.3 Proposed Models

The paper suggests two Autoencoder based DL models to detect accidents and event-based video summarization (i) Proposed Convolutional Variational Autoencoder-based framework (CVAE-ADS) and (ii) Proposed Conv-BiLSTM Autoencoder based Framework. Both of the models are created to learn normal traffic patterns and detect anomalies through deviations of learned representations. The identified abnormal frames are also used to create succinct and informative video summaries.

4.3.1 Proposed CVAE-ADS: Accident Detection and Summarization.

The suggested CVAE-ADS (Convolutional Variational Autoencoder with Anomaly Detection and Summarization) framework is aimed at identifying traffic accidents and creating summaries of events in the form of brief videos in an unsupervised fashion. Fig. 4 gives the overall architecture, and the framework works as follows in a step-by-step manner.

- **Preprocessing:** Video frames obtained in the surveillance stream are resized to 128x128 pixels and brought to the range [0,1]. This guarantees that the inputs of various camera sources are uniformly scaled and that model convergence is quicker and more robust throughout training.

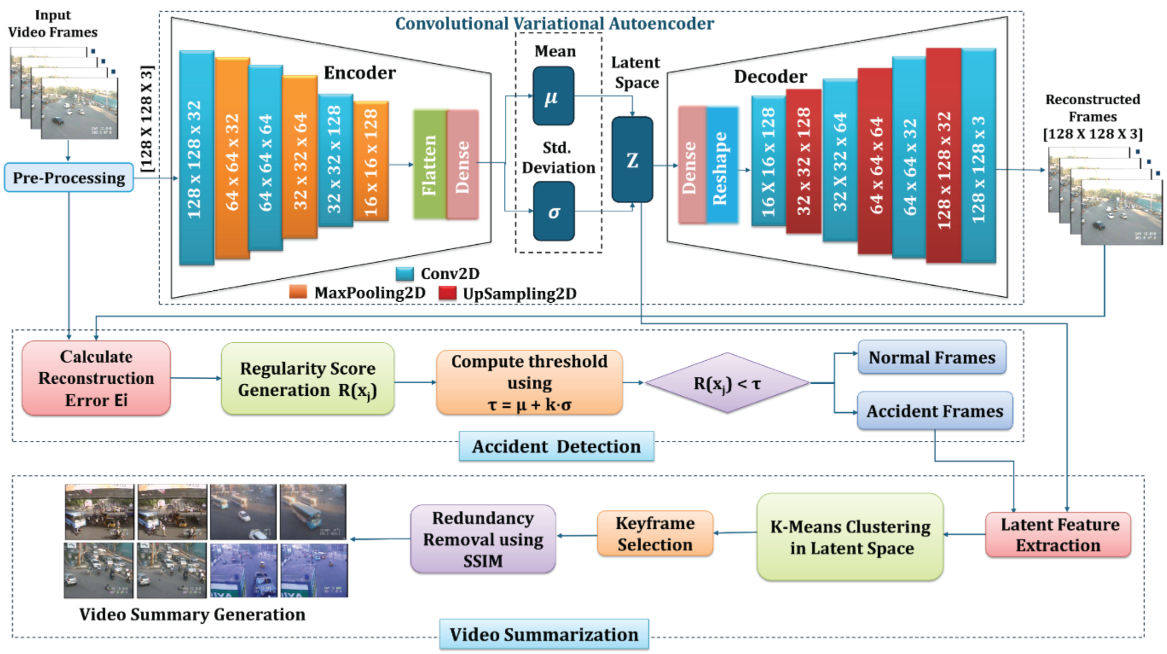


Fig. 4 Detailed architecture of the proposed CVAE-ADS based accident detection and video summarization.

- **Encoder:** The encoder is made up of layers of convolutional and pooling along with layers that extract hierarchical spatial features and map each frame to a probabilistic latent space with mean (μ) and standard deviation (σ), which enables the model to study the patterns of normal traffic distribution.

- **Latent Sampling:** The reparameterization trick is used to sample a latent vector:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

which supports stochastic representation learning and is differentiable.

- **Decoder:** The decoder assembles the input frame $\hat{x}$ of the latent vector via up sampling layers. The model is trained based on a combined objective:

$$\mathcal{L}_{CVAE} = \mathcal{L}_{recon} + \mathcal{L}_{KL} \tag{2}$$

  where the reconstruction loss makes sure that the reconstruction is correct, and the KL divergence makes the latent space regular.

- **Anomaly Detection:** In the process of inference, reconstruction error is calculated based on Mean Squared Error (MSE):

$$E(x_j) = \frac{1}{HWC}\sum(x_j - \hat{x}_j)^2 \tag{3}$$

  The error is normalized into a regularity score:

$$R(x_j) = 1 - \frac{E(x_j) - \min(E)}{\max(E) - \min(E) + \epsilon} \tag{4}$$

  A threshold τ is used to classify frames into normal or anomalous frame.

$$\tau = \mu_{err} + k\sigma_{err} \tag{5}$$

  Frames with low regularity scores $\left(R(x_j) < \tau\right)$ are identified as accident frames.

- **Video Summarization:** The identification of abnormal frames is followed by the extraction of its latent feature vectors and clustering with K-Means to form a group of accident event patterns. A representative keyframe is selected from each cluster. To eliminate visual redundancy, SSIM-based similarity filtering is applied, removing frames that are too similar to already selected keyframes. The final keyframes are arranged in chronological order to produce a compact, temporally coherent video summary that captures only the critical accident events from the original footage.

For summarization, the detected abnormal frames are encoded into latent features and grouped using K-Means clustering. Representative keyframes are selected from each cluster, followed by redundancy removal using SSIM. The final summary is generated by arranging keyframes in temporal order, providing a compact representation of accident events.

4.3.2 Proposed Conv-BiLSTM Autoencoder Framework Accident Detection and Video Summarization

The proposed Conv-BiLSTM autoencoder captures both spatial appearance and temporal dynamics of traffic video sequences for effective accident detection and event-based summarization. The overall framework and architecture are shown in Fig. 5 and Fig. 6.

- **Encoder (Spatial Features):** The encoder uses convolutional layers applied in a time-distributed manner to extract spatial features from each frame in the sequence using filters (32, 64, 128) with ReLU activation and max-pooling. The feature maps are flattened and transformed into a time sequence.

- **BiLSTM Bottleneck (Temporal Modeling):** The feature sequence is passed through stacked BiLSTM layers (128 units), which learn forward and backward temporal dependencies. This produces a compact spatio-temporal latent representation of the sequence.

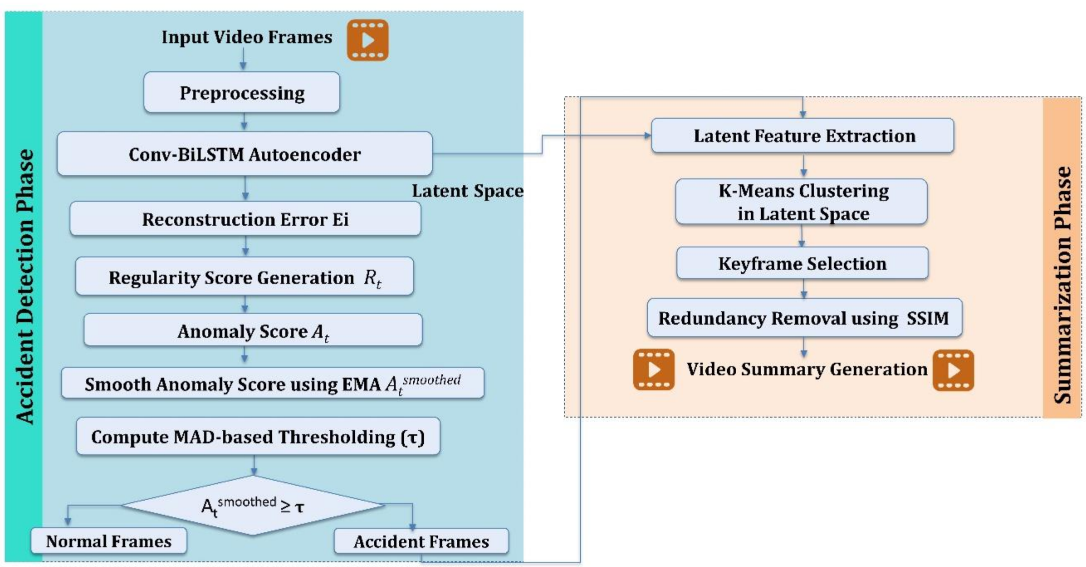


Fig. 5 Block Diagram of Proposed Conv-BiLSTM Autoencoder Framework for Accident Detection and Video Summarization

- **Decoder (Reconstruction):** The latent representation is decoded using a BiLSTM layer followed by dense and reshape operations. Transposed convolution layers progressively upsample the features, and a final sigmoid layer reconstructs the input frames.

- **Anomaly Detection:** The model is trained using normal sequences and it allows it to reconstruct regular patterns well. In the testing phase, reconstruction error (MSE) is calculated and normalized to get a regularity score. An anomaly score is determined as:

$$A_t = 1 - R_t \tag{6}$$

To stabilize the results, the scores are smoothed with the help of Exponential Moving Average (EMA). The Median Absolute Deviation (MAD) is used to calculate a strong threshold:

$$\tau = m + k \cdot MAD \tag{7}$$

Sequences that satisfy $A_t^{smoothed} \geq \tau$ are called accident events.

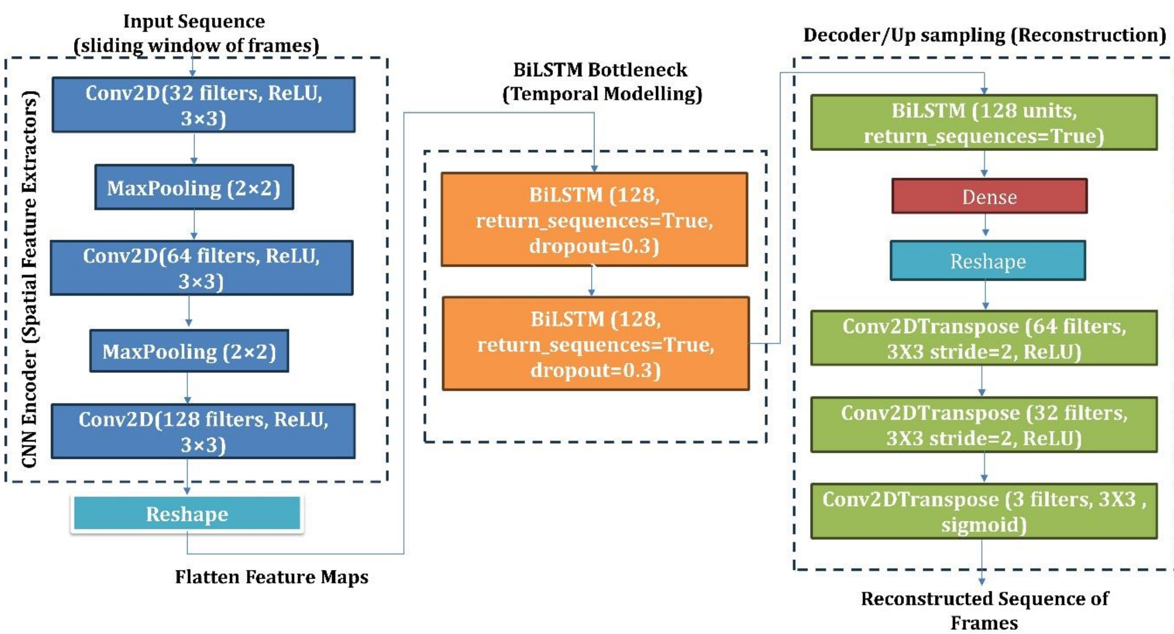


Fig. 6 Architectural diagram of Proposed Conv-BiLSTM Autoencoder based model.

- **Video Summarization:** K-Means is used to cluster similar events by clustering latent features of detected anomalous sequences. Each cluster is sampled with representative keyframes, and a filtering with SSIM eliminates redundant frames. The last summary is created by chronologically organizing the keyframes, which creates a succinct and informative presentation of the accident events.

4.4 Algorithm

In this section, the general workflow of the suggested framework on accident detection and event-based video summarization with the help of CVAE-ADS and Conv-BiLSTM models is summarized.

Input: Surveillance video V

Output: Accident frames and summarized video S detected.

1. Extract frames $F = \{f_1, f_2, \ldots, f_n\}$ from input video.
2. Normalize and resize frames so that they can be processed uniformly.
3. Generate input representation:
   - CVAE: process individual frames
   - Conv-BiLSTM: form frame sequences using sliding window
4. The model is trained on regular patterns with just normal traffic data.
5. For each test frame/sequence:
   - Reconstruct input using trained model
   - Compute reconstruction error
6. Transform reconstruction error to regularity/ anomaly score.
7. Use thresholding to determine whether frame/sequences are normal or accident:
   - CVAE: statistical threshold (mean + standard deviation)
   - Conv-BiLSTM: MAD-based threshold with temporal smoothing
8. Extract latent features of detected abnormal frames/sequences.
9. Apply K-Means clustering to group similar events.
10. Select representative keyframes from each cluster.
11. Remove redundant frames using SSIM-based filtering.
12. Arrange selected keyframes in chronological order.
13. Generate final summarized video S.

4.5 Explainability

Even with traffic surveillance applications, precise detection is not adequate; the system must be able to generate evidence that can be interpreted to back their decisions. In order to achieve transparency and reliability, the frameworks proposed will involve several levels of explainability that relate the model outputs to intuitive visual and analytical understanding.

- **Reconstruction Error Analysis:**

  Both models are learned on regular traffic patterns and anomaly detection is a reconstruction-based objective. The mistake is expressed as a regularity score graphically displayed against the video time scale. The rapid changes in this score are the signs of abnormal events, and it is easily possible to define when accidents happen and to what extent they do not correspond to normal behavior.

- **Anomaly Score Visualization:**

The complement of regularity score, which is known as an anomaly score, gives a frame-wise probability of abnormality. This score is then further smoothed with Exponential Moving Average (EMA) in the Conv-BiLSTM model as a means to reduce noise and enhance stability. Anomaly curve spikes indicate possible accident segments, which can be easily confirmed through time.

- **Latent Space Interpretation:**

In order to analyze feature separability, the latent features learned by the models are projected into a lower-dimensional space with the t-SNE. The clear separation of normal and abnormal samples indicates the existence of meaningful and discriminative representations in the models, validating the events that are detected.

- **Keyframe-based Explanation:**

The last summarized video is a human interpretable and direct output. The system can supply a concise account of the incident by displaying representative keyframes of identified events in chronological order to enable a fast comprehension of the flow of events without re-watching the entire video. All in all, the framework provides quantitative (error and anomaly scores) and qualitative (visual summaries) explanations which makes the detecting process transparent, interpretable, and applicable to the real world.

5. Results and Comparative Analysis

In this chapter, the authors provide a thorough comparison of the suggested framework on the accident detection and summarization tasks. The results of the experiments emphasize the usefulness of the suggested models, and they are further discussed in terms of comparison with the existing models to prove their strength and usefulness in practice.

5.1 Performance Evaluation

The effectiveness of the proposed approach is considered in two complementary aspects: accident detection and video summarization. As the goal of this work is not only to properly identify the accident events but to create small and informative summaries, all these are evaluated separately with a proper evaluation measure. To detect accidents, the quality of the suggested models is examined based on traditional performance measures, including accuracy, recall, precision, F1-score, and other appropriate measures. These measures aid in the comprehension of the effectiveness of the models in differentiating normal and accident frames across different traffic conditions [42] [47].

In the case of video summarization, the performance measure is the capacity of the framework to minimize redundancy and maintain critical events information. These measures include reduction rate, representativeness of the selected keyframes and the overall quality of the summary to determine how well the generated summaries represent significant accidents events [53] [54]. This two-fold assessment will guarantee a thorough analysis of the suggested solution, which will prove its detectability and its applicability in generating valuable and succinct video summaries.

5.1.1 Accident Detection Performance

- **Performance on IITH**

Table 4 and Fig. 7 present the accident detection results on the IITH Dataset. The proposed CVAE-ADS framework achieves an accuracy of 92.5%, with a precision of 0.885, recall of 0.820, F1-score of 0.851, and AUC of 0.912, outperforming the Baseline CAE (86.78%) and Baseline VAE (83.42%) by a notable margin. The Conv-BiLSTM model further improves performance, reaching 94.0% accuracy, recall of 0.91, AUC of 0.946, and a reduced EER of 14.97%, the lowest among all compared models.

Table 4 Accident Detection Performance on IITH

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC	EER (%)
Baseline CAE	86.78	0.8355	0.6740	0.7680	0.8695	24.70
Baseline VAE	83.42	0.8021	0.6020	0.7370	0.8326	26.84
Proposed CVAE-ADS	92.5	0.885	0.820	0.851	0.912	16.5
Proposed Conv-BiLSTM	94.0	0.84	0.91	0.87	0.9460	14.97

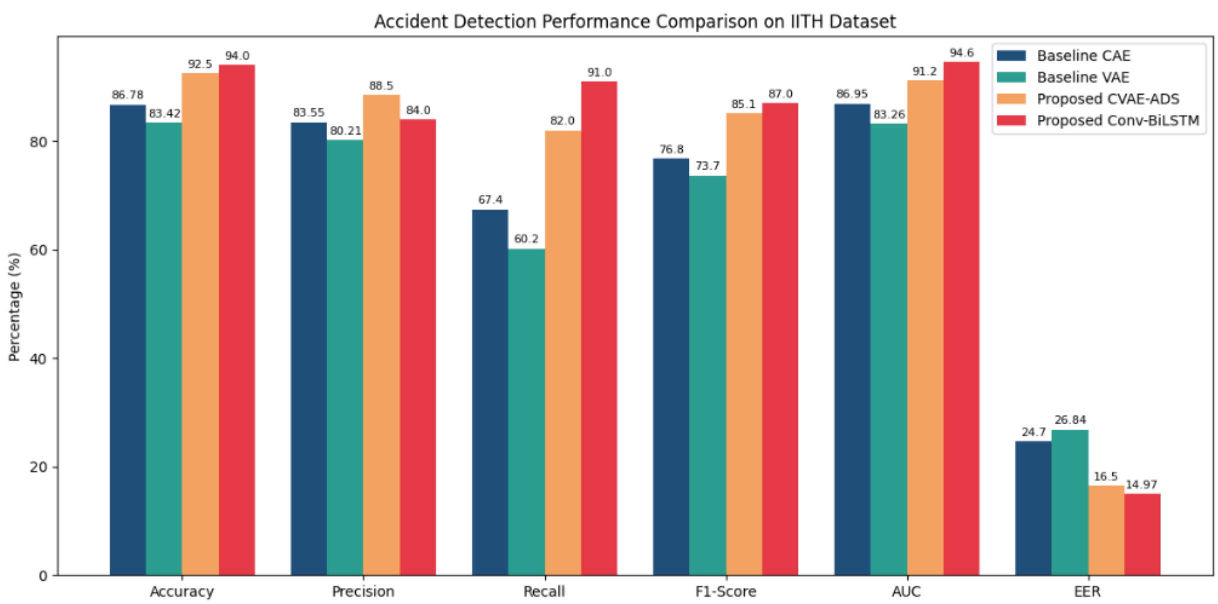


Fig. 7 Performance Comparison of Accident Detection on IITH Dataset.

- **Performance on UCF-Crime**

As shown in Table 5 and Fig. 8, similar trends are observed on the more challenging UCF-Crime Dataset. The proposed CVAE-ADS achieves 88.6% accuracy and an AUC of 0.893, while the Conv-BiLSTM delivers the best overall performance with 90.0% accuracy, recall of 0.85, AUC of 0.918, and EER of 16.20%. Both proposed models significantly outperform the baselines, which struggle with lower recall values indicating missed accident detections.

Table 5 Accident Detection Performance on UCF-Crime

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC	EER (%)
Baseline CAE	81.63	0.8012	0.6520	0.7110	0.8448	25.62
Baseline VAE	78.56	0.7658	0.6300	0.6518	0.7989	29.73
Proposed CVAE-ADS	88.6	0.872	0.791	0.829	0.893	18.7

Proposed Conv-BiLSTM	90.0	0.89	0.85	0.86	0.9180	16.20

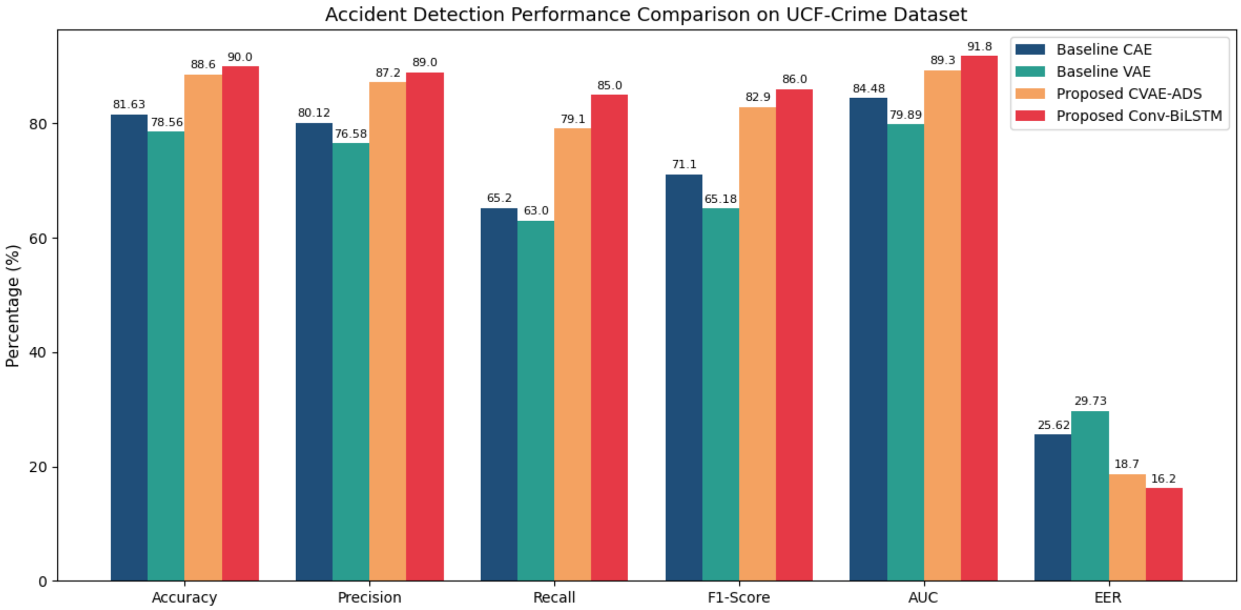
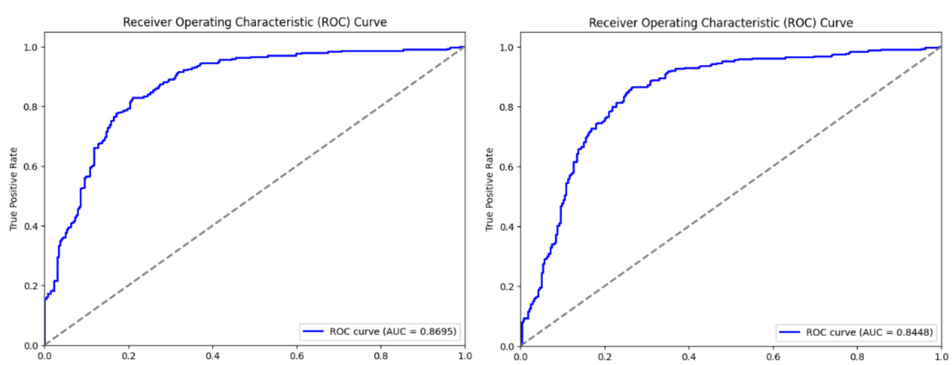


Fig. 8 Performance Comparison of Accident Detection on UCF-Crime Dataset.

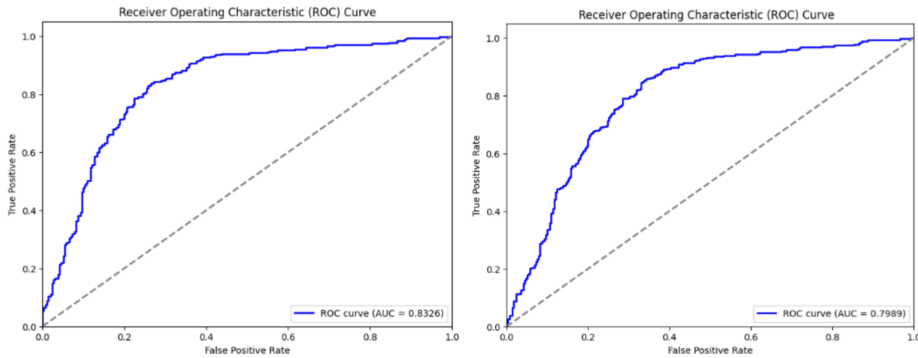
- **ROC Curve Analysis**

Fig. 9 and Fig. 10 indicate that the ROC curves of the baseline models show moderate discrimination (AUC: 0.7989-0.8695). Conversely, Fig. 11 and Fig. 12 indicate that the proposed models have higher AUC values with CVAE-ADS having AUCs of 0.912 and 0.893 and Conv-BiLSTM having AUCs of 0.946 and 0.918 on both the datasets, respectively. This means that it is more likely to detect and less false positive.



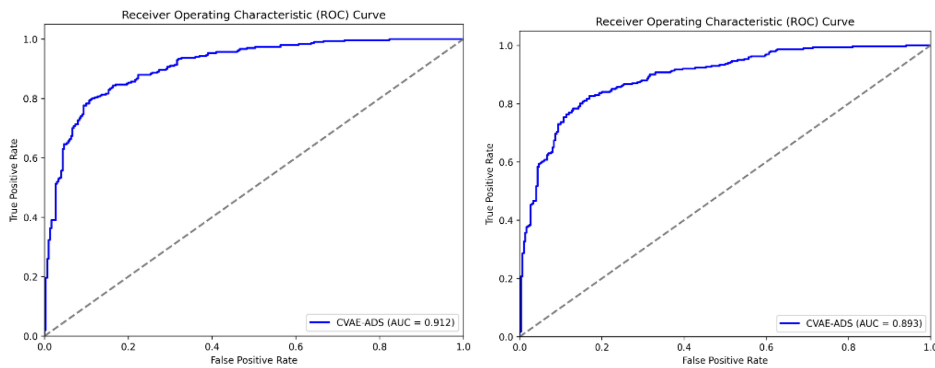
(a) (b)

Fig. 9 ROC curves for Baseline CAE Model: (a) the IIT Hyderabad and (b) UCF-Crime Dataset



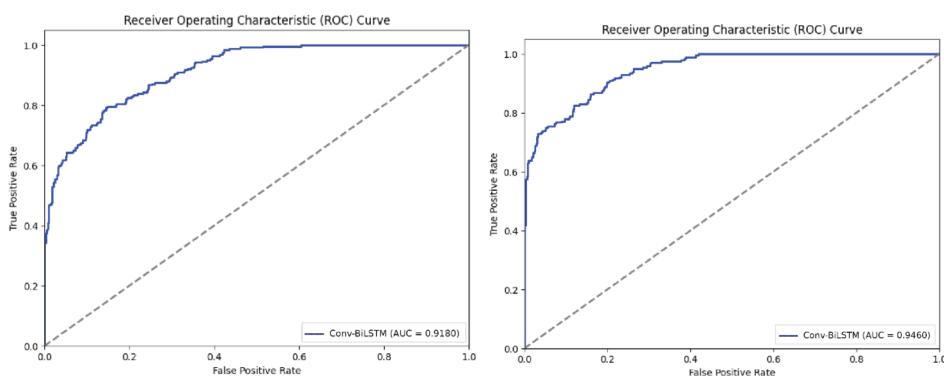
(a) (b)

Fig. 10 ROC curves for Baseline VAE Model: (a) the IIT Hyderabad and (b) UCF-Crime Dataset



(a) (b)

Fig. 11 ROC curves for CVAE-ADS Model: (a) the IIT Hyderabad and (b) UCF-Crime Dataset



(a) (b)

Fig. 12 ROC curves for Conv-BILSTM Model: (a) the IIT Hyderabad and (b) UCF-Crime Dataset

In both datasets, the Conv-BiLSTM model achieves better results than both baselines, and the CVAE-ADS, due to its capability to learn temporal dynamics using two-way LSTM layers, and

Table 6 CVAE-ADS Video Summarization Performance

Metric	IIT Hyderabad	UCF-Crime Accident
Reduction Rate (%)	70-85%	70–80%
Diversity Rate (1 − SSIM)	0.7249	0.70
Silhouette Score (t-SNE)	0.56	0.52
PSNR	30.1 dB	28.8 dB

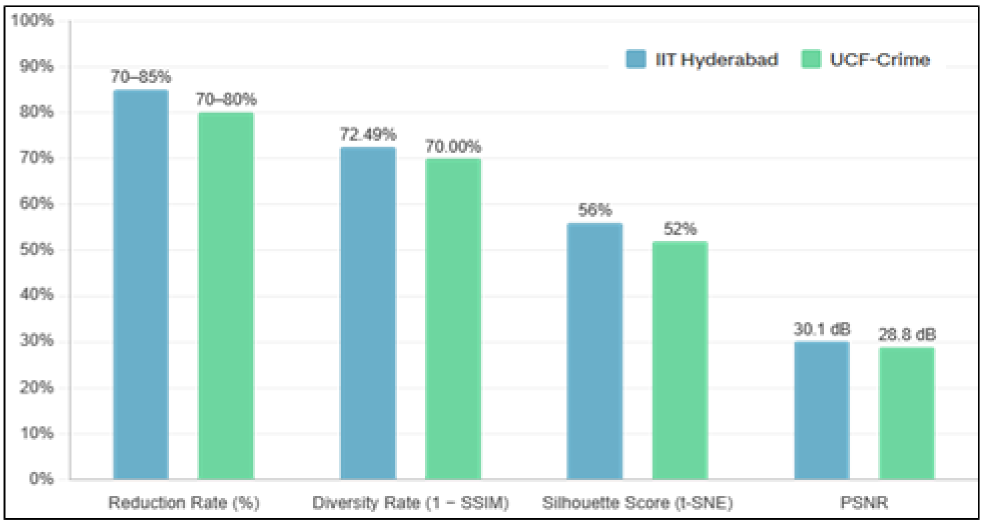


Fig. 13 Performance Comparison of CVAE-ADS Summarization Model.

spatial features. CVAE-ADS, in its turn, is an option that provides a good compromise between detection and computational complexity and can be considered an option in resource-limited deployments. Both models show that the reconstruction-error based unsupervised anomaly detection is a highly viable and feasible tool in real-world traffic monitoring.

5.1.2 Video Summarization

This section will assess the performance of the proposed models in creating small and informative video summaries of the IITH and UCF-Crime datasets. Reduction Rate, Diversity Rate (1-SSIM), Silhouette Score, and PSNR are some of the metrics used to analyze the performance.

- **Performance of CVAE-ADS Model**

The proposed CVAE-ADS model has a high reduction rate of 70-85% on IITH dataset and 70-80% on UCF-Crime dataset as shown in Table 6 and Fig. 13, which means it is effective at eliminating redundant frames. The diversity scores (0.7249 and 0.70) indicate that the summarized frames are still diverse enough. Also, the silhouette scores (0.56 and 0.52) show that there is a reasonable clustering of latent features, and PSNR values of 30.1 dB and 28.8 dB show that the visual quality of the reconstructed summaries is acceptable.

- **Performance of Conv-BiLSTM Model**

The performance of the Conv-BiLSTM-based summarization is presented in Table 7 and Fig. 14. The model has a similar reduction rate (6585) but has a higher diversity (0.8567 and 0.82) and better silhouette scores (0.64 and 0.55) on both datasets. Also, it can be seen that PSNR values are higher (32.45 dB and 31.3 dB), which is an indicator of a higher quality of reconstruction, which means that it is able to retain significant visual information.

Table 7 Conv-BiLSTM-based Video Summarization Performance

Metric	IIT Hyderabad	UCF-Crime Accident
Reduction Rate (%)	70-85%	65–80%
Diversity Rate (1 − SSIM)	0.8567	0.82
Silhouette Score (t-SNE)	0.64	0.55
PSNR	32.45	31.3

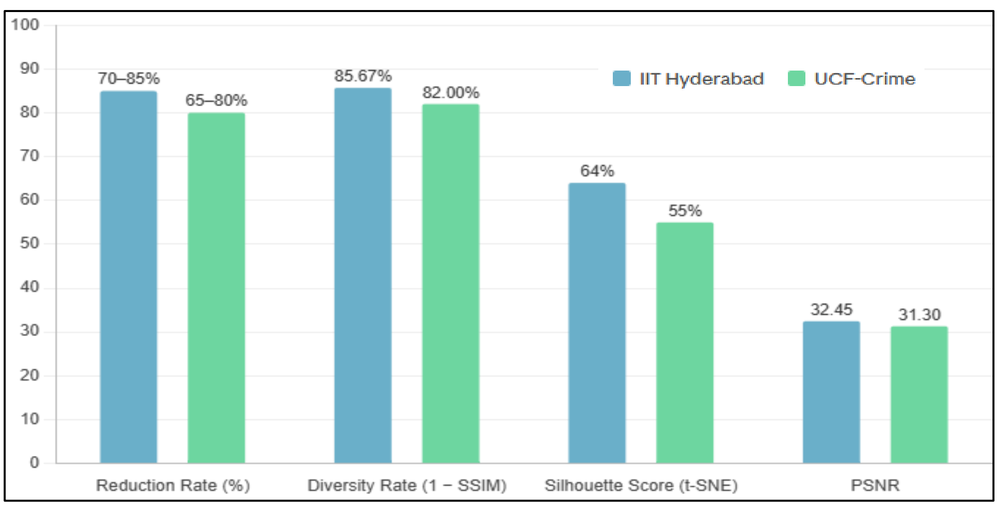
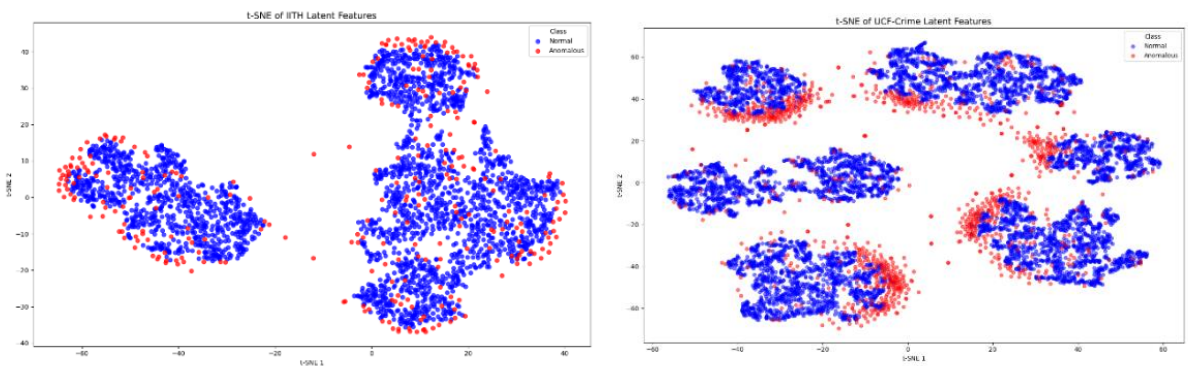


Fig. 14 Performance Comparison of Conv-BILSTM Model.

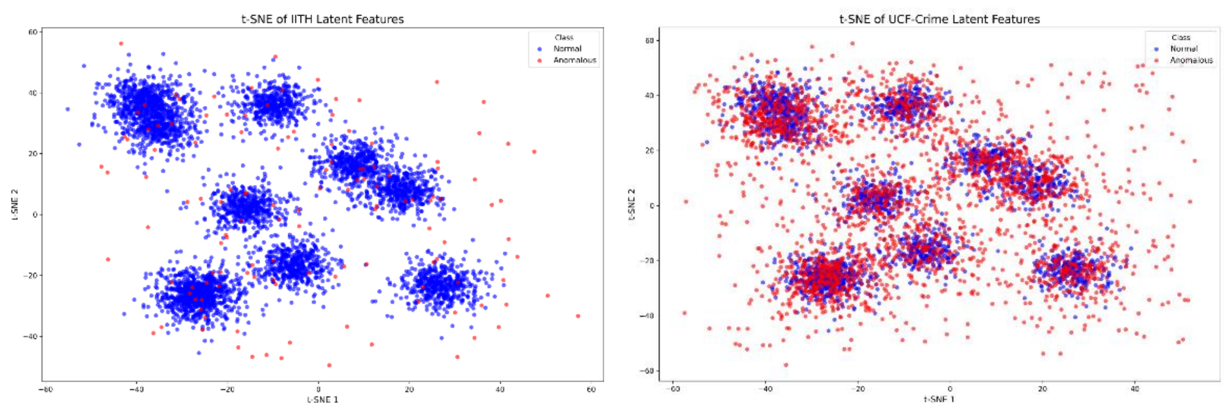
- **Latent Space Analysis**

The t-SNE plots in Fig. 14 and Fig. 16 show the latent features distribution of both models. The clusters between normal and anomalous frames are moderately separable with the CVAE-ADS model, and more compact and well-defined with the Conv-BiLSTM model, which suggests that the model features a better representation of features and better ability to distinguish between important events.



(a) (b)

Fig. 15 t-SNE of Latent Features for the proposed CVAE-ADS Model: (a) IITH Dataset and (b) UCF-Crime.



(a)

(b)

Fig. 16 t-SNE of Latent Features for the proposed Conv-BiLSTM Model: (a) IITH Dataset and (b) UCF-Crime.

To demonstrate the practical effectiveness of the proposed approaches, a sample-based evaluation is conducted on selected videos from both the datasets. It is important to note that only a subset of videos is presented, and therefore, this analysis provides indicative insights rather than a complete comparison.

Tables 8 and 9 present the performance of the CVAE-ADS approach on sample IITH and UCF-Crime videos. For the IITH dataset, the model retains most accident frames while achieving 71%–77% summarization reduction. For example, in Video 1, 240 out of 305 frames are retained with 77.1% reduction and 4.20× compression. For the UCF-Crime dataset, similar trends are observed, with reduction values ranging from 73% to 79%. For instance, Video 2 retains 278 frames with 78.6% reduction and 4.68× compression. On the whole, CVAE-ADS has a desirable balance among accident frame retention and compression, but a few frames are lost.

Table 8 CVAE-ADS: Performance Evaluation on Sample IITH Videos

Video	Original Frames	Accident Frames	Retained (TP)	Missed (FN)	Summarized Frames	Summarization Reduction (%)	Compression (×)
Video 1	2,941	305	240	65	70	77.1	4.20
Video 2	2,432	104	82	22	28	73.1	4.05
Video 3	4,200	85	66	19	22	74.1	3.95
Video 4	1,439	53	43	10	16	71.7	3.70
Video 5	2,980	225	187	38	52	76.8	4.35

Table 9 CVAE-ADS: Performance Evaluation on Sample UCF-Crime Videos

Video	Original Frames	Accident Frames	Retained (TP)	Missed (FN)	Summarized Frames	Summarization Reduction (%)	Compression (×)
Video 1	918	201	159	42	50	75.1	4.02
Video 2	1,773	351	278	61	75	78.6	4.68
Video 3	4,449	151	119	32	40	73.5	3.77
Video 4	1,799	151	125	26	38	74.8	3.97
Video 5	2,580	351	278	56	72	79.5	4.87

Tables 10 and 11 give a summary of the CONV-BiLSTM model performance. In the case of the IITH data, the model is more effective in reducing (77%- 82) and retention is enhanced. As an example, Video 1 has 250 frames and a reduction of 80.3 and a compression of 5.08. In the case of UCF-Crime dataset, the reduction is between 77 and 81 and the data is similar across videos. For instance, Video 5 achieves 80.9% reduction and 4.15× compression. All in all, CONV-BiLSTM offers a higher reduction and maximum compression with retaining competitive accident frames in comparison with CVAE-ADS.

Table 10 CONV-BiLSTM: Performance Evaluation on Sample IITH Videos

Video	Original Frames	Accident Frames	Retained (TP)	Missed (FN)	Summarized Frames	Summarization Reduction (%)	Compression (×)
Video 1	2,941	305	250	55	60	80.3	5.08
Video 2	2,432	104	85	19	22	78.8	4.73
Video 3	4,200	85	70	15	18	78.8	3.89
Video 4	1,439	53	43	10	12	77.4	3.58
Video 5	2,980	225	185	40	40	82.2	4.63

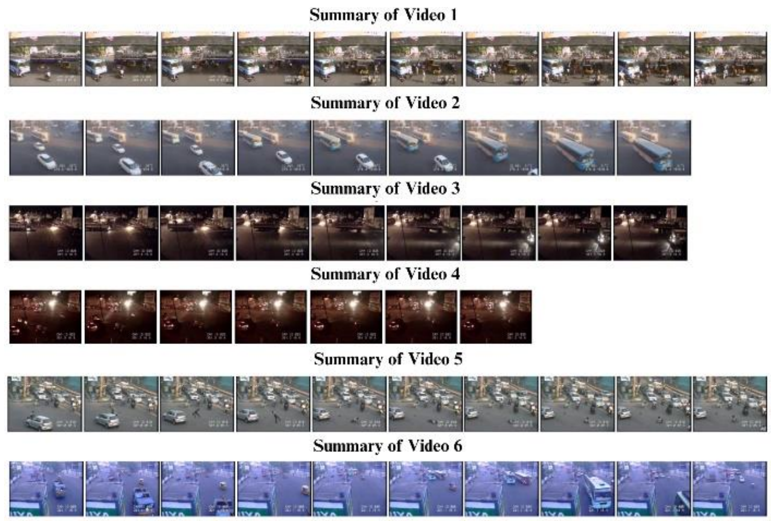
Table 11 CONV-BiLSTM: Experiment Results on Sample UCF-Crime Videos.

Video	Original Frames	Accident Frames	Retained (TP)	Missed (FN)	Summarized Frames	Summarization Reduction (%)	Compression (×)
Video 1	918	201	159	42	45	77.6	3.53
Video 2	1,773	351	278	73	70	80.1	3.97
Video 3	4,449	151	119	32	32	78.8	3.72
Video 4	1,799	151	119	32	31	79.5	3.84
Video 5	2,580	351	278	73	67	80.9	4.15

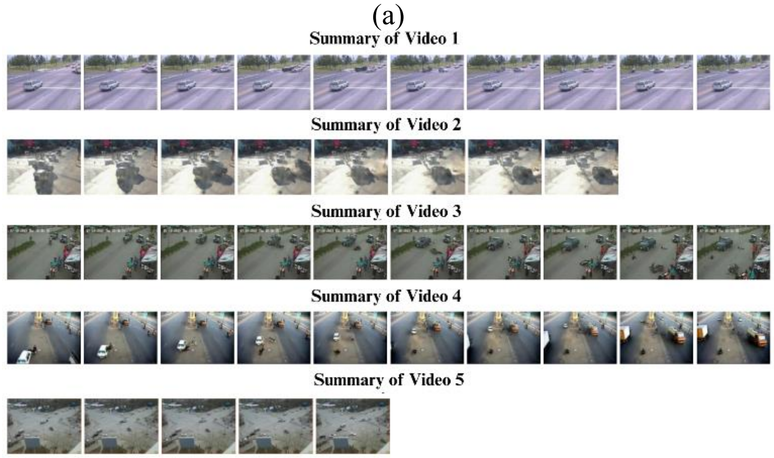
The qualitative summarization of the sample videos in the IITH and UCF-Crime datasets with the suggested models are discussed in Figure 17 and 18. It is necessary to note that there are few videos demonstrated to be visually illustrated.

CVAE-ADS model demonstrates that the model is able to capture important moments related to accidents and eliminate redundant frames. The resulting summaries are concise, and concentrate on key events, but there are instances where some irrelevant frames are not discarded, suggesting a balanced but somewhat conservative summarization behaviour.

The model Conv-BiLSTM provides better and more refined summaries whereby the identified frames are more closely localized to accident occurrences. Model generates less and more accurate summaries, minimizing redundancy but preserving significant information on time. Comprehensively, the two models have effective balances between compression and information retention. Although CVAE-ADS is more efficient in summarization with less complexity, Conv-BiLSTM model offers better diversity, quality of clustering, and reconstruction. This underscores the benefit of using temopral modeling to produce more informative and credible video summaries. The two models have been successful in outlining meaningful accident segments. Nevertheless, the CONV-BiLSTM model yields rather succinct and narrower summaries, and CVAE-ADS offers consistent and stable coverage of events.

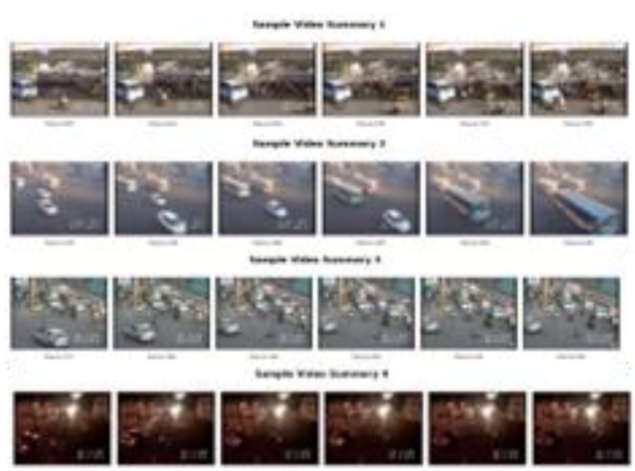


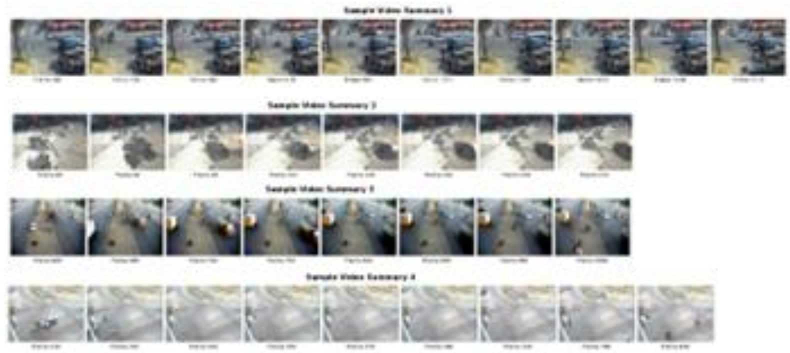
(a)



(b)

Fig. 17 Result of Sample video summarization by the proposed CVAE-ADS: (a) IITH and (b) UCF-Crime.





(b)
Fig. 18 Result of Sample video summarization the proposed Conv-BILSTM: (a) of IITH (b) UCF-Crime

5.2 Comparative Analysis

The relative study on the existing state-of-the-art techniques, presented in Table 12, helps to demonstrate that the suggested methods are effective in various benchmarks. Previous studies like D. Singh et al. [42] and Pawar et al. [47] use unsupervised methods of autoencoders, however, with much lower AUC (7779) and larger EER. Supervised algorithms, such as Wang et al. [45] and Srinivasan et al. [44], achieve better detection rates (92.5% at most) and AUC (96.32%), but are based on both labelled data and complicated structures. Conversely, the unsupervised but competitive CVAE-ADS model obtains an AUC of 91.2% (IITH) and 89.3% (UCF-Crime) as well as lower EER values. Moreover, the suggested Conv-BiLSTM model shows better performance with AUC of 94.60% (IITH) and 91.80% (UCF-Crime) and lower EER (14.97% and 16.20%), which implies a better detection rate and strength. The proposed models offer an improved trade-off between performance and practicality as compared to the previous approaches, especially in terms of dependency on labeled data and high detection accuracy.

Table 12 Comparison with State-of-the-Arts Accident Detection.

Ref.	Learning	Approach	Dataset	AUC	EER	mAP	Detection Rate	FAR
[42]	Unsupervised	Unsupervised Denoising Autoencoder + One-Class SVM	IITH	77%	22.50%	-	-	-
[44]	Supervised	DETR + Random Forest		82%	-	-	78.2%	-

[45]	Supervised	Retinex + YOLOv3 + Decision Tree	Online CCTV videos	96.32%	-	-	92.5%	7.5%
[46]	Supervised	CNN + Rolling Prediction	VAID & Test Dataset	-	-	-	88%	-
[47]	Unsupervised	Conv-AE + Seq2Seq LSTM Autoencoder	IITH	79%	20.50%	60%	-	-
			DOTA	84.70%	11.7%	-	-	-
[48]	Supervised	YOLOv2 + Transfer Learning	IITH	-	-	76%	-	-
[49]	Semi-Supervised	Object Interaction + Refinement + Heatmaps	UCF Crime	69.70%	-	-	-	0.8
			CADP	72.59%	-	-	-	2.2
[50]	Supervised	Lightweight I3D - ConvLSTM2D	Custom Dataset	-	-	87%	80%	-
Our Work	Unsupervised	Proposed CVAE-ADS	IITH	91.2%	16.5%	-	-	-
			UCF-Crime	89.3%	18.7%	-	-	-
		Proposed Conv-BiLSTM Approach	IITH	94.60%	14.97%			
			UCF-Crime	91.80%	16.20%			

Table 13 Comparison of state-of-the-art video summarization techniques.

Ref.	Approach / Model	Dataset(s) Used	Reduction Rate (%)	Diversity Rate	PSNR (dB)
[51]	Perceptual Video Summarization	YouTube-8M, Urban Tracker	Used saliency	×	×
[52]	Video Synopsis using ResNet	IITH	×	×	×
[53]	Z-STRFG: Spatio-temporal fuzzy approach	YouTube8M, Anomaly20, Urban Tracker	×	×	×
[54]	YOLOv5 + Privacy-preserved summarization	Synthetic + Real-time Data	42.97%	×	×

[55]	YOLOv5 + Event-based summarization	Synthetic + Real Traffic Videos	20–50%	×	×
Proposed CVAE-ADS Approach	Conv-VAE + Latent Space Clustering	IITH,	70-85%	0.7249	30.1
		UCF-Crime	70-80%	0.70	28.8
Proposed Conv-BiLSTM Approach	Conv-BiLSTM + Latent Space Clustering	IITH	70-85%	85.67	32.45
		UCF-Crime	65–80%	0.82	31.3

The comparison to the existing video summarization methods, as illustrated in Table 13, shows that the proposed methods are quite efficient in reducing, diverse, and being visually appealing. Previous algorithms like those of Thomas et al. [51], Kosambia et al. [52], and Pramanik et al. [53], are mainly concerned with perceptual or spatio-temporal approaches but do not provide quantitative values such as reduction rate, diversity or PSNR. More current methods based on YOLOv5 (Mehwish Tahir et al. [54] and Saxena et al. [55]) have moderate reduction rates (around 2050 percent), but do not include diversity assessment and quality evaluation. Conversely, the CVAE-ADS solution is much more effective in reducing the image (70-85 and 70-80 percent, respectively) and retains good diversity (0.72 and 0.70) and reasonable visual quality (PSNR up to 30.1 dB). In addition, the suggested Conv-BiLSTM method demonstrates superior performance with a high reduction rate (up to 85%), high diversity (up to 0.82), and high reconstruction quality (PSNR up to 32.45 dB). In general, the proposed models are superior to current approaches, offering a more balanced trade-off among compression, diversity, and visual quality, and are, therefore, more applicable to practice in video summarization.

6. Achievements with Respect to Objectives

The major accomplishments of this study are as follows and prove how the objectives set have been met successfully:

1. **Development of a Unified Framework for Accident Detection and Video Summarization**

   • **Achievement:** A combination of CVAE-ADS and Conv-BiLSTM Autoencoder models has been successfully built into a unified deep learning framework. The architecture collectively conducts accident detection and video summarization based on reconstruction-based anomaly learning and latent space clustering without relying on any labelled data.

   • **Outcome:** Experimental results show that the proposed system is effective at detecting anomalous (accident) frames and producing concise summaries, a gap that has been filled by focusing on detecting and summarizing as distinct tasks.

2. **Effective Learning of Spatio-Temporal Features**

   • **Achievement:** CVAE-ADS model has the advantage of capturing spatial representations, whereas the Conv-BiLSTM autoencoder improves the temporal sequence modeling with the help of the bidirectional dependencies.

   • **Outcome:** The hybrid strategy enhances perception of motions and scene dynamics, resulting in better detection of accidents and contextualized summarization.

3. **Reconstruction-Based Anomaly Detection**

   • **Achievement:** Accident events are estimated with reconstruction error thresholds, i.e., when the patterns of normality are deviated, the abnormality is detected.

   • **Outcome:** This facilitates dependable detection in an unsupervised environment, and avoids the need to rely on annotated datasets, while maintaining high detection rates.

4. **Generation of Compact and Informative Video Summaries**

   • **Achievement:** A keyframe selection strategy based on latent space-clustering is employed to acquire representative frames of observed events.

   • **Outcome:** The framework tends to reduce video by a substantial amount (up to ~85%), maintaining critical information about accidents in good visual quality.

5. **Improved Performance and Efficiency**

• **Achievement:** The proposed models yield better results compared to baseline methods (CAE, VAE): accuracy, AUC, F1-score, and EER reduction, but at the same time, they are computationally efficient.

• **Outcome:** The system offers optimal trade-off between detection accuracy, compression efficiency and reconstruction quality, and can be used in real-world surveillance.

7. Conclusion and Future Work

This work presents a unified and unsupervised framework for accident detection and video summarization CVAE-ADS and Conv-BiLSTM model based on autoencoders. The suggested solution overcomes the main weaknesses of the current solutions since it combined detection and summarization in a single pipeline, hence enhancing efficiency and consistency. The CVAE-ADS model is effective in spatial representations and anomaly detection through reconstruction error, and the Conv-BiLSTM autoencoder is efficient in improving the temporal modeling through sequence dependencies in both directions.

Latent space modeling, combined with clustering, makes it possible to choose representative keyframes, which provide informative summaries that are compact. In experiments conducted on the IIT Hyderabad and UCF-Crime Accident subset datasets, the proposed models are shown to be more accurate, more precise, more recalling, higher F1-score, and higher at the AUC compared to the baseline methods, and lower in the Equal Error Rate (EER). The framework also reduces video by a significant amount (up to approximately 85 percent) without important information regarding the accident being lost as well as it has a high reconstruction quality with PSNR values exceeding 30 dB.

In general, the proposed system offers a compromise between the accuracy of detection, the efficiency of summarization, and the quality of the visuals, which is why it can be used in the context of intelligent surveillance in the real world. Specifically, Conv-BiLSTM model is more temporally consistent and produces more accurate summaries than the CVAE-ADS method.

Future Work:

- Attention mechanisms and transformer-based architecture can play a crucial role in advancing the model with the capability to concentrate on the important spatial regions and learn long-range temporal correlations across video sequences. This can enhance the knowledge on complex patterns of accidents that change with time and result in more accurate detection and generalization.
- The framework can be extended to incorporate multi-modal data, e.g. audio cues, vehicle sensor data or contextual cues can be used to complement the information, thus enhancing the strength and reliability of accident detection, particularly in difficult or uncertain conditions.

- It is necessary to optimize the proposed models in order to use them in real-time. This will be done using model compression methods like pruning, quantization, and lightweight architectural design, which will allow it to run efficiently in resource constrained surveillance settings.
- Researching self-supervised and semi-supervised learning methods can decrease the reliance on particular data distributions and the number of labeled data. This will make the framework more flexible to the unknown environments and different traffic conditions.
- A step further to apply the analysis to large and varied real-world data sets, such as varying weather some time, camera angles and traffic congestion will assist to prove the generalizability and robustness of the proposed system in field applications.

Publications

1. Chauhan, A., Vegad, S. (2022). Smart Surveillance Based on Video Summarization: A Comprehensive Review, Issues, and Challenges. In: Suma, V., Fernando, X., Du, KL., Wang, H. (eds) Evolutionary Computing and Mobile Sustainable Networks. Lecture Notes on Data Engineering and Communications Technologies, vol 116. Springer, Singapore. https://doi.org/10.1007/978-981-16-9605-3_29
2. A. Chauhan and S. Vegad, "CVAE-ADS: A Deep Learning Framework for Traffic Accident Detection and Video Summarization", j.electron.electromedical.eng.med.inform, vol. 8, no. 1, pp. 185-205, Jan. 2026. 10.35882/jeeemi. v8i1.1139 (A Scopus Indexed Journal)
3. Ankita Chauhan and Dr. Sudhir Vegad, Trans., "Road Traffic Accident Detection Using Autoencoder-Based Models: A Performance Analysis of CAE and VAE", Int J Sci Res Sci & Technol, vol. 13, no. 2, pp. 298–313, Mar. 2026, Doi: 10.32628/IJSRST261330. (Peer Reviewed Journal)

References

1. Sathyamoorthy, M., Rajasekar, V., Krishnamoorthi, S., & Pamucar, D. (2026). Ensemble deep learning approach for traffic video analytics in edge computing. *Scientific Reports*.
2. Vrochidis, A., Panagiotidis, S., Parcharidis, S., Dimitriou, N., Papaioannou, P., Milolidaki, E., & Krinidis, S. (2025). Leveraging 5G networks for event video summarization via multimodal analysis. *Discover Artificial Intelligence*.
3. Noor, A., Almukhalfi, H., Noor, T. H., & Ranjan, R. (2026). TAERM: Traffic Accident Emergency Response Management Framework for Detection and Classification Using IoT and YOLOv9. *Future Generation Computer Systems*, 108444.
4. Li, X., Ye, X., & Sun, Q. (2024). Leveraging Connected Vehicle Data for Near-Crash Detection and Analysis in Urban Environments. *arXiv preprint arXiv:2409.11341*.
5. Almutairi, A., Al Asmari, A. F., Alanazi, F., Alqubaysi, T., & Armghan, A. (2025). Deep learning based predictive models for real time accident prevention in autonomous vehicle networks. *Scientific Reports*, *15*(1), 20844.
6. Khan, A. N., Shah, Y. A., Khan, W., Khalil, A., & Khan, J. (2025). A meta fusion model combining geographic data and twitter sentiment analysis for predicting accident severity. *Scientific Reports*, *15*(1), 25122.
7. Deshpande, U. U., Shanbhag, S., Patil, R., Chate, R. A. A., Das Chagas Silva Araujo, S., Pinto, K., ... & Nirwani, B. (2025). Automatic two-wheeler rider identification and triple-riding detection in surveillance systems using deep-learning models. *Discover Artificial Intelligence*, *5*(1), 104.
8. Zhou, W., Yang, L., Zhao, L., Zhang, R., Cui, Y., Huang, H., ... & Wang, C. (2025). Vision technologies with applications in traffic surveillance systems: A holistic survey. *ACM Computing Surveys*, *58*(3), 1-47.
9. Li, W., Huang, L., & Lai, X. (2025). A deep learning framework for traffic accident detection based on improved YOLO11. *Vehicles*, *7*(3), 81.
10. Alzamzami, O., Alsaggaf, Z., AlMalki, R., Alghamdi, R., Babour, A., & Al Khuzayem, L. (2025). Passable: An Intelligent Traffic Light System with Integrated Incident Detection and Vehicle Alerting. *Sensors*, *25*(18), 5760.
11. Sun, W., Abdullah, L. N., Khalid, F. B., & Suhaiza Binti Sulaiman, P. (2025). EAR-CCPM-Net: A Cross-Modal Collaborative Perception Network for Early Accident Risk

Prediction. *Applied Sciences*, *15*(17), 9299.

12. Kim, H., Khudoyberdiev, A., SR Garnaik, S., Bhattacharya, A., & Ryoo, J. (2025). CLOUD-CODEC: A New Way of Storing Traffic Camera Footage at Scale. *ACM Transactions on Multimedia Computing, Communications and Applications*, *21*(8), 1-28.
13. Huang, Z., Pan, G., Kang, C., & Lv, Y. (2025). Semi-supervised pipe video temporal defect interval localization. *Computer-Aided Civil and Infrastructure Engineering*, *40*(17), 2418-2433.
14. Bhosale, M. K., Patil, S. B., & Patil, B. B. (2025). Automatic Video Traffic Surveillance System with Number Plate Character Recognition Using Hybrid Optimization-Based YOLOv3 and Improved CNN. *International Journal of Image and Graphics*, *25*(04), 2550041.
15. Ismail, A., Ismail, A. R., Shaharuddin, N. A., Ara, M. A., Puzi, A. A., Awang, S., & Ramli, R. (2025). Vision-based vehicle classification for smart city. *Aptisi Transactions on Technopreneurship (ATT)*, *7*(2), 441-453.
16. Singh, H., Verma, M., & Cheruku, R. (2025). Hsnet: A novel edge-preserving hierarchical separable network for video shadow detection. *Circuits, Systems, and Signal Processing*, *44*(6), 3983-4012.
17. Khanam, S., Sharif, M., Raza, M., Ishaq, W., Fayyaz, M., & Kadry, S. (2025). Anomaly recognition in surveillance based on feature optimizer using deep learning. *PLoS One*, *20*(5), e0313692.
18. Peng, D., & Yan, W. (2025). Test-time training with adaptive memory for traffic accident severity prediction. *Computers*, *14*(5), 186.
19. Madhumitha, G., & Senthilnathan, R. (2025). Development of a deep learning image classification network based on vehicular collision using instance mask guided attention. *Signal, Image and Video Processing*, *19*(4), 342.
20. Arefin, M. S., Mahin, M. I. S., & Mily, F. A. (2025). Real-time rapid accident detection for optimizing road safety in Bangladesh. *Heliyon*, *11*(4).
21. Srinivasan, K. (2025). Object tracking using optimized Dual interactive Wasserstein generative adversarial network from surveillance video. *Knowledge-Based Systems*, *311*, 113084.
22. Shihab, M. A., Hussein, M. K., Abd, N. S., & Ahmed, S. R. (2025). Towards Safer Roads: An Automated Traffic Control System with Enhanced Vehicle Speed Detection and

Measurement Algorithm. *International Journal of Computing and Digital Systems*, *17*(1), 1-17.

23. Xia, Y., Qian, N., Guo, L., & Cai, Z. (2024). CF-SOLT: Real-time and accurate traffic accident detection using correlation filter-based tracking. *Image and Vision Computing*, *152*, 105336.
24. Rani, S., & Dalal, S. (2024). Comparative analysis of machine learning techniques for enhanced vehicle tracking and analysis. *Transportation Engineering*, *18*, 100271.
25. Ahmed, A., Farhan, M., Eesaar, H., Chong, K. T., & Tayara, H. (2024). From detection to action: A multimodal AI framework for traffic incident response. *Drones*, *8*(12), 741.
26. Liang, C., Chen, Q., Li, Q., Wang, Q., Zhao, K., Tu, J., & Jafaripournimchahi, A. (2024). HADNet: A novel lightweight approach for abnormal sound detection on highway based on 1D convolutional neural network and Multi-Head Self-Attention mechanism. *Electronics*, *13*(21), 4229.
27. Wu, X., & Li, T. (2024). A deep learning-based car accident detection approach in video-based traffic surveillance. *Journal of Optics*, *53*(4), 3383-3391.
28. Bhanja, U., Mohanty, A., & Mahapatra, S. (2024). Fuzzy logic-based accident detection system for vehicular ad hoc networks: a prototype implementation. *Wireless Personal Communications*, *138*(1), 291-320.
29. SP, K., Mohandas, P., & CS, S. (2024). Smart junction: advanced zone-based traffic control system with integrated anomaly detector. *Annals of operations research*, *340*(1), 479-506.
30. Zahid, A., Qasim, T., Bhatti, N., & Zia, M. (2024). A data-driven approach for road accident detection in surveillance videos. *Multimedia Tools and Applications*, *83*(6), 17217-17231.
31. Natha, S., Jokhio, F. A., Laghari, M., Siraj, M., Alsaif, S. A., Ashraf, U., & Ali, A. (2024). A scalable and generalized deep ensemble model for road anomaly detection in surveillance videos. *Comput. Mater. Contin*, *81*, 3707-3729.
32. Pai, A. A., Harini, K. S., Giridhar, D., & Rangaswamy, S. (2023). Real-Time Accident Detection from Closed Circuit Television and Suggestion of Nearest Medical Amenities. *IJITCS*, *15*(6), 15-23.
33. Nasry, A., Ezzahout, A., & Omary, F. (2023). People tracking in video surveillance systems based on artificial intelligence. *Journal of Automation, Mobile Robotics and*

Intelligent Systems, 59-68.

34. Zhang, Y., & Sung, Y. *Traffic accident detection method using trajectory tracking and influence maps, Mathematics (2023)*.
35. Alanazi, J. M. (2023). Effective Machine-Learning Based Traffic Surveillance Moving Vehicle Detection. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, *14*(1), 95-105.
36. Zhu, L., Wang, B., Yan, Y., Guo, S., & Tian, G. (2022). A novel traffic accident detection method with comprehensive traffic flow features extraction. Signal Image Video Process.
37. Basheer Ahmed, M. I., Zaghdoud, R., Ahmed, M. S., Sendi, R., Alsharif, S., Alabdulkarim, J., ... & Krishnasamy, G. (2023). A real-time computer vision based approach to detection and classification of traffic incidents. *Big data and cognitive computing*, *7*(1), 22.
38. Vijay, T. K., Dogra, D. P., Choi, H., Nam, G., & Kim, I. J. (2022). Detection of road accidents using synthetically generated multi-perspective accident videos. *IEEE Transactions on Intelligent Transportation Systems*, *24*(2), 1926-1935.
39. Tahir, M., Qiao, Y., Kanwal, N., Lee, B., & Asghar, M. N. (2023). Real-time event-driven road traffic monitoring system using CCTV video analytics. *IEEE Access*, *11*, 139097-139111.
40. Pan, H., Guan, S., Zhao, X., & Xue, Y. (2023). Edge computing-based vehicle detection in intelligent transportation systems. *Computing and Informatics*, *42*(6), 1339-1359.
41. Taheri Sarteshnizi, I., Sarvi, M., Bagloee, S. A., & Nassir, N. (2023). Temporal pattern mining of urban traffic volume data: a pairwise hybrid clustering method. *Transportmetrica B: Transport Dynamics*, *11*(1), 2185496.
42. Singh, D., & Mohan, C. K. (2019). Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. IEEE Transactions on Intelligent Transportation Systems, 20(3), 879–887. https://doi.org/10.1109/TITS.2018.2835308
43. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6479–6488).
44. Srinivasan, A. Srikanth, H. Indrajit and V. Narasimhan, "A Novel Approach for Road Accident Detection using DETR Algorithm," 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 2020, pp. 75-80,

doi:10.1109/IDSTA50958.2020.9263703.

45. Wang, Chen & Yulu, Dai & Zhou, Wei & Geng, Yifei, “A Vision- Based Video Crash Detection Framework for Mixed Traffic Flow Environment Considering Low-Visibility Condition,” Journal of Advanced Transportation, 2020. https://doi.org/10.1155/2020/9194028
46. Khan, Sardar Waqar, Qasim Hafeez, Muhammad Irfan Khalid, Roobaea Alroobaea, Saddam Hussain, Jawaid Iqbal, Jasem Almotiri, and Syed Sajid Ullah. 2022. "Anomaly Detection in Traffic Surveillance Videos Using Deep Learning" Sensors 22, no. 17: 6563. https://doi.org/10.3390/s22176563
47. Karishma Pawar, Vahida Attar, Deep learning based detection and localization of road accidents from traffic surveillance videos, ICT Express, 2021, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2021.11.004.
48. A. R. Pathak and A. C. Elster, "Applying Transfer Learning to Traffic Surveillance Videos for Accident Detection," 2022 International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway, 2022, pp. 1-7, doi: 10.1109/ICAPAI55158.2022.9801568.
49. K. V. Thakare, D. P. Dogra, H. Choi, H. Kim and I. -J. Kim, "Object Interaction-Based Localization and Description of Road Accident Events Using Deep Learning," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 20601-20613, Nov. 2022, doi: 10.1109/TITS.2022.3170648.
50. V. A. Adewopo and N. Elsayed, "Smart City Transportation: Deep Learning Ensemble Approach for Traffic Accident Detection," in IEEE Access, vol. 12, pp. 59134-59147, 2024, doi: 10.1109/ACCESS.2024.3387972.
51. S. S. Thomas, S. Gupta and V. K. Subramanian, "Event Detection on Roads Using Perceptual Video Summarization," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 9, pp. 2944-2954, Sept. 2018, doi: 10.1109/TITS.2017.2769719.
52. Kosambia, Twinkle and Gheewala, Jaydeep, Video Synopsis for Accident Detection using Deep Learning Technique (May 22, 2021). Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Available at SSRN: https://ssrn.com/abstract=3851250 or http://dx.doi.org/10.2139/ssrn.3851250
53. A. Pramanik, S. K. Pal, J. Maiti and P. Mitra, "Traffic Anomaly Detection and Video Summarization Using Spatio-Temporal Rough Fuzzy Granulation With Z-Numbers," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 24116-

24125, Dec. 2022, doi: 10.1109/TITS.2022.3198595.

54. Tahir M, Qiao Y, Kanwal N, Lee B, Asghar MN. Privacy Preserved Video Summarization of Road Traffic Events for IoT Smart Cities. Cryptography. 2023; 7(1):7. https://doi.org/10.3390/cryptography7010007
55. Saxena, N., Asghar, M.N. (2023). YOLOv5 for Road Events Based Video Summarization. In: Arai, K. (eds) Intelligent Computing. SAI 2023. Lecture Notes in Networks and Systems, vol 739. Springer, Cham. https://doi.org/10.1007/978-3-031-37963-5_69