Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 1, January 2026, pp: 185-205 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

CVAE-ADS: A Deep Learning Framework for Traffic Accident Detection and Video Summarization

Ankita Chauhan[1] and Sudhir Vegad[2]

[1] Department of Computer/ IT Engineering, Gujarat Technological University, Ahmedabad, India
[2] Department of Computer Engineering, G H Patel College of Engineering & Technology, The Charutar Vidya Mandal (CVM) University, Vallabh Vidhyanagar, India

Corresponding author: Dr. Sudhir Vegad (e-mail: Sudhir.vegad@cvmu.edu.in) **Author(s) Email**: Ankita Chauhan. (e-mail: ankita.chauhan@cvmu.edu.in)

Abstract **Since it is a manual process of monitoring to identify accidents, it is becoming more and more difficult and results in human error, because of the rapid increase in road traffic and surveillance video. This underscores the urgent need for robust, automated systems capable of identifying accidents, as well as the burden of summarizing long videos. In order to address this issue, we propose CVAE-ADS, which is an unsupervised Approach that not only detects anomalies but also summarizes keyframes of a video to monitor traffic. This method operates in two phases. The stage of detecting Abnormalities intraffic video is performed using a Convolutional Variational Autoencoder, which operates on normal frames and identifies anomalies based on reconstruction errors. The second stage is the clustering of the perceived anomalous frames in the latent space, followed by the selection of representative keyframes to form a summary video. We tested the method with two benchmark datasets, namely, the IITH Accident Dataset and a subset of UCF-Crime. The findings have shown that the proposed approach had great accuracy of accident detection and AUC of 90.61 and 87.95 on IITH and UCF-Crime respectively and low rebuilding error and Equal Error Rates. To summarize, the method achieves substantial frame reduction and produces high visual quality with a wide variety of keyframes. It is able to measure up to 85 reduction rates with coverage of 92.5 on the IITH dataset and 80 reduction rates with coverage of 90 on an Accident subset of the UCF-Crime Dataset. CVAE-ADS offers a lightweight version of constant traffic monitoring, which utilizes limited organizational capital to categorize coincidences in real-time and recapitulate video footage of the accidents.**

Keywords **Anomaly Detection, Video Summarization, Convolutional Variational Autoencoder, Latent Space Clustering**

I. Introduction

The World Health Organization (WHO) report documented that 20-50 million damages or incapacities per year and about 1.19 million deaths are caused by traffic accidents. More than 90 per cent of road traffic deaths are registered in low and middle-income countries, and the lowest percentage in Europe. The death of individuals between the ages of 5 and 29 as a result of traffic injuries in the majority of cases. These collisions also have a huge financial cost, accounting for approximately 3 percent of the GDP of countries in lost productivity, medical expenses, and care costs. To address this, one of the United Nations stipulations is to reduce road traffic deaths and injuries by 50 percent by 2030 [1]. The United Nations has set a target to minimize these deaths and injuries by half in 2030. According to the Ministry of Road Transport and Highways (MoRTH) report, 4,61,312 accidents were reported in 2022, resulting in 4,43,366 serious injuries and 1,68,491 deaths, higher than the previous year [2] [3]. The majority of road accidents were a consequence of violation of traffic regulations, of which over-speeding caused 1,19,904 deaths. The other significant violations included wrong-side driving, use of a mobile phone while driving, and driving under the influence of alcohol, and other violations, which further emphasize the significance of the violation of traffic regulations on the road. It is also worth mentioning that other roads had the highest numbers of accidents with 39.4 percent, National Highways with 36.2 percent, and State Highways with 24.3 percent, hence the need to implement an alternative or efficient accident detection system [2]. As computer vision and DL methods have developed, Intelligent Transportation Systems (ITS) have brought about the booming development of video processing and given an urgent need to evaluate abnormal situations in video [4]. Surveillance videos provide anomalous event detection, which is regarded as the most complicated and difficult issue in the

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

context of computer vision and DL, with numerous areas of application in surveillance [5]. The masses of CCTV photographic cameras currently being developed in the public and private areas create the necessity of intelligent video monitoring. In this study, we have taken the idea of coincidence discovery as an anomaly detection problem, in which case accidents are viewed as an exceptional deviation of normal traffic patterns. Typically, conventional methods relying on handcrafted characteristics are not resistant and generalizable since lighting conditions or movement differ [10]. DL based models, including CNNs, LSTMs, GANs, and even autoencoders, are more effective at handling the problem of these conditions. Rather than manually designed features, it is able to automatically detect patterns in the data and address them effectively. It will render them more adaptive when dealing with various datasets.

Most existing accident detection systems [13-16] [18-20] do not integrate detection and synopsis, which creates an overreliance on supervised algorithms and large-scale annotated datasets. It is very hard to acquire and annotate traffic accident data due to the unpredictability, the rarity, and the diverse environmental conditions in which traffic accidents occur. Supervised learning models trained on limited, scene-specific datasets often do not generalise to new traffic situations or to the photographic camera angles covered in the scenes they were trained on. Conversely, unsupervised methods learn normal traffic behaviour without any labelling information and detect abnormal behaviour as a deviation of the resulting model, providing enhanced scalability, flexibility, and usability in continuous surveillance scenarios where annotated accident videos are limited or unavailable. This poses challenges for real-time environment deployment, particularly in contexts with limited annotated accident datasets. Additionally, numerous approaches focus solely on detection. They do not offer succinct visual summaries for quick incident evaluation, reducing their practical usefulness for traffic management centres.

Identifying abnormal accident incidents and the Condensation of significant frames are both very critical issues in traffic surveillance systems. It minimizes the use of manual video inspections, thereby enhancing quicker and more precise automatic identification of incidents, thereby refining road safety and emergency responses [8]. Due to the growing necessity of video Recapitulation, numerous approaches have been suggested to meet the numerous applications and supplies of the different learning techniques, including supervised, semi-supervised, unsupervised, and reinforcement learning [6] [9]. The vsLSTM [25] and the attention-based BiLSTM [27] are supervised learning methods that use labelled data to select the most meaningful frames. None of the techniques uses the less exact process of summarizing, although recent techniques implement spatial-temporal modelling and object tracking for the more specific task: DHAVS [28], and the YOLOv5-DeepSORT-SSD pipeline [29]. An example of such a model, like YOLOv8 [49] and YOLOv11 [50], is created to diagnose an accident with the help of a video recording of real traffic. Unsupervised techniques, in their turn, do not rely on labelled data. Instead, they are trained by statistical regularities or deep features, including object-level summarization using sparse LSTM autoencoders [30] or GAN-based models, including CNN Bi-ConvLSTM-GAN [33]. Semi-supervised methods are a compromise between the two methods, involving both labelled and unlabelled data, e.g., VESD [36] and hierarchical RL systems, such as VDAN+ [37], to achieve an efficient sum-up at a lower cost. Reinforcement learning models are still described as sequential tasks based on rewards and optimizing on diversity and informativeness, e.g., Deep Summarization Network (DSN) [38], 3D Spatio-temporal U-Net [40], and PRLVS [41]. All these strategies emphasize the vitality of video summarization in dynamic and multifaceted substantial.

Convolutional Variational Autoencoder (CVAE) has been adopted in this rather than an ordinary autoencoder or the traditional CNN due to the principal nature of the task of detecting accidents in surveillance video, which relies on the ability to model normal patterns and identify deviations of these patterns through reconstruction error [10, 42, 44]. In contrast to a traditional autoencoder, a VAE is trained to generate a probabilistic distribution of normal traffic patterns, continuity, and structure in the latent space are imposed through Kullback-Leibler divergence [42, 43]. Such property is necessary for anomaly detection, which allows better separation between rare abnormal events (accidents) and the learned normal behavior [45, 47, 48]. Additionally, integrating convolutional layers into the VAE preserves the spatial structure of a traffic scene, including lane structures, vehicle positions, and distortions caused by impacts [12]. Such spatial characteristics are normally sacrificed by fully connected autoencoder structures [10]. As a result, when an accident occurs, the CVAE produces significantly higher reconstruction errors, and thus, the reliability of anomaly scoring is increased [42, 44, 45]. Moreover, the structured latent space obtained by the CVAE provides a useful embedding, which can be further used to perform K-Means clustering, which underpins the proposed key-frame selection and summarization procedure [22, 30, 32].

This study aims to create and evaluate a lightweight, unsupervised framework that can detect traffic accidents and produce brief visual summaries in real-

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

time from surveillance videos, eliminating the requirement for manual annotations.

In this research, the proposed approach is designed and evaluated using real-world traffic surveillance videos obtained from fixed CCTV cameras installed at highways and urban intersections. The data sets used have diverse conditions such as variant lighting (day lighting, night lighting, early morning), traffic congestion, changing weather, and partial occlusions, hence, realistic traffic conditions. The current execution assumes a comparatively fixed camera angle and makes no specific allowances regarding dynamically tracking a photographic camera, aerial shots, or extreme unfavorable conditions, like heavy fog or storm events. Such restrictions identify possible ways of improving and expanding the proposed framework in the future. Specifically, the primary contributions of the given paper in the Accident detection and video summarization include:

1. Convolutional-VAE Accident Discovery and Summarizer (CVAE-ADS) is a complete system that we are proposing to detect accidents based on reconstruction-driven anomaly scoring using a single convolutional variational autoencoder that is trained solely on normal traffic scenes.
2. To decrease redundancy and emphasize important content, we fit a latent space clustering-based outline plan, which is effective in choosing keyframes of accidents as representatives to review the incident effectively.
3. Our unsupervised learning-based approach requires no manual labelling, hence it is scalable for large surveillance networks at low cost.
4. Moreover, the design is tailored for real-world oriented architecture and easily integrated into the traffic monitoring systems, which facilitate post-processing tasks like emergency response, insurance verification, and legal accountability.

The presented CVAE-ADS framework is feasible in the real world. Its lightweight, unsupervised nature enables integration into a real-time traffic monitoring system, enabling timely accident detection and concise summary generation without manual data labeling. The capabilities are particularly useful in traffic-intensive care facilities, emergency response, and insurance and litigation appraisal, as well as post-processing near-miss incidents, where rapid and precise interpretation is essential. The following outlines the structure of the current article. Section II provides a complete analysis of related work, including traffic accident detection, video summarization, and the basic principles of Variational Autoencoders for anomaly detection. Section III offers a comprehensive description of the proposed CVAE-ADS framework that encompasses architectural design, a two-stage processing pipeline, model components, and hyperparameter specifications. It also describes the datasets, evaluation metrics, and experimental set up used to evaluate accident detection and summarization. In Section IV, we will report and analyze the experimental results, including the accuracy of accident detection, the summarization effectiveness, the visual analyses, and the comparison against state-groups. Ultimately, Section V looks at some important observations, strengths, and limitations of the approach studied in this research paper. Meanwhile, Section VI tries to conclude the study and gives directions to future research and enhancements to this system.

II. Related Work

Various authors have proposed various approaches and techniques for recognizing or classifying accidents. In the related work, a review of various accident detection and video Summarization approaches is discussed. While video anomaly detection has been extensively addressed in literature, few works consider road-accident detection and video summarization in a single unified unsupervised framework. Lack of real-world accident data for aviation, privacy concerns in surveillance, and high cost of annotating detailed data are all factors that hinder access to large datasets with task-specific focus. For this reason, this section offers a brief yet comprehensive review of works related to accident detection and video summarization, as direct or indirect anomaly detection issues in traffic situations.

A. Related Work on Accident Detection

Singh & Mohan [12] introduced an unsupervised method using Stacked Denoising Autoencoders and one-class SVM for accident detection, achieving 77.5% accuracy. Though effective in varied lighting, it struggles with occlusions, night scenes, and traffic complexity. Srinivasan et al. [13] proposed a road accident detection method combining DETR for object detection and a Random Forest Classifier for event prediction, achieving a 78.2% detection rate. While effective, DETR shows limitations in detecting small or low-visibility objects. Wang et al. [14] developed a vision-based crash detection framework for low-visibility traffic scenes using Retinex for image enhancement, YOLOv3 for object detection, and a decision tree for classification. It achieved 92.5% accuracy with a 7.5% false alarm rate. Robles-Serrano et al. [15] suggested a DL architecture that uses CNN and LSTM to learn spatiotemporal features of accidents in traffic videos. Although it works well in diverse conditions, its performance declines in scenes of high traffic density. Khan et al. [16] applied CNN and rolling prediction in detecting the presence of an accident anomaly, where 82 percent was obtained, but the results decreased in foggy or remote scenes. Pawar et

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

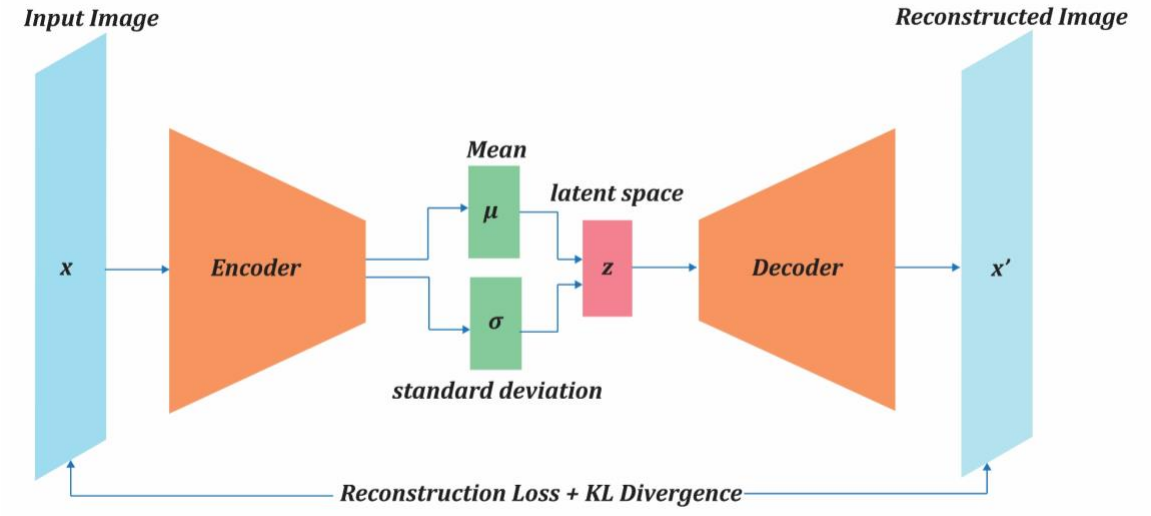


Fig. 1. An Architecture of a Variational Autoencoder

al. [17] suggested a DL-based model in which they integrated a Convolutional Autoencoder and a Seq2Seq LSTM to describe spatiotemporal features of the accident. It scored 79% and 84.7% on the IITH dataset and DOTA, respectively, which is 11.7 times higher than that of any other unsupervised method. Pathak & Elster [18] proposed an accident detection model using YOLOv2 with transfer learning, trained on the UCF-Crime dataset and tested on the IITH dataset, achieving 76% mAP.

An object interaction-based approach for accident detection, localization, and severity description, utilizing heatmaps and textual summaries, is proposed by Thakare et al. [19]. It achieved AUCs of 69.7% (UCF-Crime) and 72.59% (CADP), with low false alarm rates, showing competitive performance. Adewopo & Elsayed et al. [20] introduced a lightweight I3D-CONVLSTM2D model that fuses RGB and optical flow for accident detection in smart city traffic videos, achieving 87% mAP and outperforming existing methods. While these techniques have improved accuracy using DL and object detection pipelines, there are still some issues. Many techniques rely on explicit object detection, handcrafted motion features, optical flow features, or predefined thresholds, which typically perform poorly in complex real-world scenarios such as occlusion, high vehicular density, low-lighting conditions, poor visibility due to adverse weather, and moving cameras. Furthermore, most techniques consider crash detection as a standalone activity and lack a holistic way to condense long-duration videos. This makes them less useful for real-time traffic surveillance and automatic crash detection in intelligent transportation systems.

B. Related Work on Road Accident Video Summarization

Research on road accident detection and summarization remains relatively limited, with most existing work focusing on general video summarization or traffic anomaly detection using traditional vision methods or DL models like CNNs. This article surveys a select set of studies specifically addressing accident-focused video summarization. Thomas et al. [7] developed a perceptual video summarization method for accident detection using the YouTube-8M and Urban Tracker datasets, showing strong performance across metrics like saliency cost and detection rate. But it could only handle single-camera shots and was not able to differentiate collisions and near-collisions, which implies the necessity of depth-based segmentation. The article by Kosambia et al. [21] applied a video synopsis method to the IITH dataset to detect accidents with a ResNet, and ResNet-152 performed best. They suggested parameter tuning, additional training data, and localization techniques to further enhance performance. Pramanik et al. [22] proposed Z-STRFG, a Z-number and spatio-temporal rough fuzzy granulation-based system on the YouTube8M, Urban Tracker, and Anomaly20 to detect anomalies and video summaries of traffic videos. It was more accurate and fast in the complex conditions, but was confined to monocular vision and had no multimodal support, which impacted the real-time applicability.

Tahir et al. [23] proposed a privacy-preserving video summarization approach for accident detection using YOLOv5, trained on a synthetic dataset and tested on real-time data with 55–85% accuracy. Privacy is ensured through synthetic training data and video encryption, with summaries reducing duration by an average of 42.97%. Saxena et al. [24] employed YOLOv5 for accident detection and proposed event-based video summarization to reduce surveillance data

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

storage. Trained on synthetic data and tested on real traffic videos, the approach effectively detected accidents and cut video length by 20–50%, improving storage efficiency.

Although these studies demonstrate the need for summarization regarding video length and storage, it should also be noted that resolving object detection confidence, manual heuristics, and synthetic training data are all problems that must be addressed. These limitations stand to affect all generalizations involving uncontrolled, real world traffic systems. More than this, there are no systems that depend on the latent structure of the video to generate representative summaries coherently. Most of all, there is simply no attempt to present a separate, unified, and unsupervised architecture where accident detection and summary are treated as a singular, unified problem.

C. Overview of Variational Autoencoder

Variational Autoencoders (VAEs) are generative models that extend traditional autoencoders by encoding input data into a probabilistic latent space, enabling the generation of realistic and diverse samples. It is particularly effective for image generation and data compression, as well as in anomaly detection [42] [43]. Fundamentally, VAEs are based on deep neural networks, but they employ a Bayesian inference principle. A latent representation learned by VAEs is structured and continuous, which ensures that points in the latent space produce meaningful outputs when decoded. This allows VAEs to generate new data points by sampling them from the learned latent space [44].

Fig. 1 shows the architecture of a variational autoencoder, which consists of three key components: an encoder, a latent space, and a decoder. The encoder compresses input data into a latent representation by learning some parameters that define a probability distribution, usually a Gaussian. VAEs do not sample from these distributions directly. It uses a reparameterization trick that makes an entire model differentiable and thus trainable via backpropagation. A sample is drawn from the latent space and passed to the decoder, which attempts to reconstruct the original input. The training process involves minimizing two losses: one that ensures the output closely matches the input, and another that regularizes the latent space to follow a known distribution, often a standard normal. This architecture enables VAEs to learn meaningful, structured data representations that are useful for generating new data samples and for anomaly detection.VAEs used in anomaly detection (particularly in tasks related to accident detection) are trained on normal traffic scenes only. They get to know how to recreate them. Nevertheless, under abnormal conditions such as accidents, the quality of reconstructions reduces, leading to an increase in reconstruction errors, which is an indication of anomalies. Learned latent features are also useful to give a summary of the scene, and these can be used to assist downstream tasks such as classification or keyframe extraction [44]-[47].

Most existing approaches come with significant challenges. Traditional models like SVMs and decision trees rely on handcrafted features and often fail in real-world traffic due to occlusion, blur, or poor lighting. DL models, while more advanced, usually need large amounts of labelled data, which is both time-consuming and expensive to collect. On top of that, many of these models are too computationally heavy for real-time use. Even though the standard VAE, which is primarily used for representation learning and generative modeling, is directly applied in advanced traffic surveillance scenarios, it remains limited. in this research, we incorporate both VAE and convolutional architectures by employing additional layers and creating the Convolutional Variational Autoencoder-based accident detection and summarization (CVAE-ADS). Unlike all other approaches, where detection and summarization are considered as independent processes, our system utilizes the learned latent distribution not only for highly efficient normal-accident pattern discrimination but also for latent clustering-based keyframe extraction, which we believe is a novel approach. This fully integrated, self-organizing system has a low annotation effort requirement, is robust under various traffic conditions, and has the ability to create efficient and meaningful summaries for quick incident analysis.

III. Methodology

As simple frame reconstruction is made possible in traditional methods such as PCA or sparse coding, it is not usually sufficient to analyze complex spatial patterns and the overlaying structure of video scenes in the real world. video scenes. Overall, such forms of linear methods are unable to receive the richer spatial context, and it is difficult to identify subtle anomalies that are not just pixel differences. Instead, our proposed CONV-VAE model utilizes deep convolutional encoders, which can learn rich multi-level features as well as a probabilistic latent space that recognizes the nuanced irregularities in a better way from the traffic videos. This approach not only improves anomaly detection but also provides meaningful video summarization that conventional approaches are not designed to handle.

A. Overview of CVAE-ADS

The proposed CVAE-ADS approach is a two-stage framework that efficiently detects accidents and summarizes key frames from traffic surveillance videos. Initially, the Convolutional Variational

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Autoencoder (CVAE) is trained only on normal traffic scenes. During the testing phase, the model reconstructs normal frames very well; but does not reconstruct the accident frames correctly, which in turns to high reconstruction errors. This error is then used for reconstruction-based anomaly scoring, where higher reconstruction error indicates a greater likelihood that the frame is anomalous (i.e., contains an accident). These values are converted into regularity **scores**, where lower scores represent a higher degree of abnormality and therefore a higher confidence of an accident occurrence.

In the second stage, the internal compressed latent features generated by the proposed CVAE from detected anomalous frames are extracted. These latent features, often referred to as the latent space, encode compact, meaningful information about the visual content of each frame. To group visually similar accident frames and remove redundancy, these latent vectors are clustered using the K-Means algorithm. This process, referred to as latent space clustering, groups frames with similar visual characteristics into distinct clusters. From each cluster, a representative frame is selected as a keyframe. These selected keyframes are then arranged chronologically, are compiled into a short video that highlights key moments of the incident, supporting quick review and analysis. Fig. 2 presents the process flow diagram of the proposed CVAE-ADS.

B. Model Architectures

The suggested model combines a Convolutional Autoencoder (CAE) with a Variational Autoencoder (VAE) with the encoder being created to remove the hierarchical spatial features of traffic frames with the help of convolutional operations and nonlinear activations and progressive down sampling. The mean and log-variance are parameterized by separate dense layers to permit sampling the latent space stochastically using the reparameterization trick. The decoder mirrors the encoder using deconvolutional and up-sampling to reconstruct the input frames. This design allows the model to effectively learn the distribution of normal traffic patterns and generate accurate reconstructions, making it suitable for unsupervised anomaly detection. Fig. 3 shows details of the internal architecture of the Convolutional Variational Autoencoder used for accident detection and video summarization.

C. Model Architectures and Hyperparameter Specifications

Table 1 shows the Hyperparameter and Architecture Specifications of the proposed CVAE-ADS. Each video frame is resized to 128 × 128 × 3 before being passed into the CVAE encoder. The encoder is composed of multiple convolutional layers (Conv2D) with increasing filter sizes (32, 64, and 128) and ReLU activation, followed by max-pooling to progressively reduce spatial dimensions while preserving essential visual information. The resulting feature maps are flattened and connected to a dense layer, from which two parallel layers estimate the mean (μ) and standard deviation (σ) of the latent distribution. A latent vector z is then sampled using the reparameterization trick.

The decoder mirrors the encoder structure and reconstructs the input frame by applying a dense layer, followed by up-sampling, and convolution operations to progressively restore the spatial resolution. The final layer uses a Sigmoid activation to generate the reconstructed frame matching the original input size. The model is trained using the Adam optimizer with a learning rate of 0.0001, a batch size of 64, and 100 training epochs. The total loss combines the mean squared reconstruction error with the Kullback–Leibler divergence, allowing the CVAE to learn a compact representation of normal traffic patterns and to detect accidents through increased reconstruction error during testing.

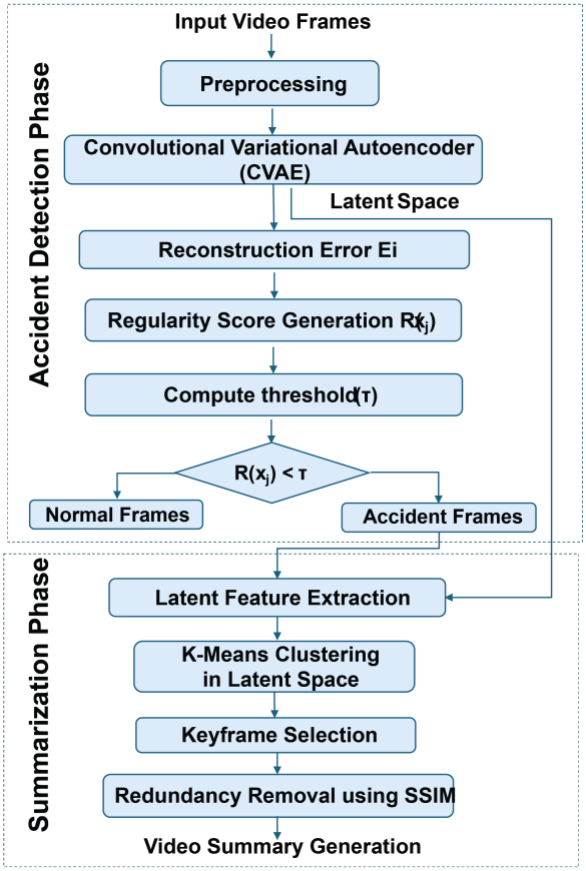


Fig. 1. Process flow diagram showing the two-stage

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Table 1. Hyperparameter and Architecture Specifications of CVAE-ADS

Component	Specification
Input size	128 × 128 × 3
Conv1	32 filters, 3×3, ReLU + MaxPooling
Conv2	64 filters, 3×3, ReLU + MaxPooling
Conv3	128 filters, 3×3, ReLU + MaxPooling
Dense (encoder)	128 units
Latent dimension (z)	64
Decoder input reshape	16 × 16 × 128
Deconv layers	128 → 64 → 32 → 3 filters
Output activation	Sigmoid
Optimizer	Adam
Learning rate	0.0001
Batch size	64
Epochs	100
Loss function	MSE + KL Divergence

D. Training Protocol and Implementation Details

The proposed CVAE-ADS framework is trained in an unsupervised manner using only normal traffic frames. Prior to training, all input frames are resized to 128 × 128 × 3 and normalized to the range [0, 1]. The model is optimized using the Adam optimizer with a learning rate of 0.0001 and a batch size of 64. Training is performed for a maximum of 100 epochs. The objective function is defined as a weighted combination of the mean squared reconstruction loss (MSE) and Kullback–Leibler (KL) divergence, where the reconstruction term penalizes pixel-level differences between the input and reconstructed frames, and the KL term regularizes the latent distribution towards a standard normal distribution. A 30% validation split is used during training to monitor generalization. To prevent overfitting and ensure stable convergence, early stopping is applied with a patience of 10 epochs, and the best-performing model weights are restored automatically. Network convergence is determined by the stabilization of both the training and validation loss curves.

E. Convolutional - VAE for Anomaly Detection

The proposed CVAE-ADS framework leverages a Convolutional Variational Autoencoder (CVAE) to detect anomalies in traffic video frames. After standard preprocessing and normalization (as described in Section D), each video frame is fed into the CVAE encoder for anomaly modelling. The encoder is a stack of multiple convolutional layers with ReLU activations that are progressively reduced in size by max-pooling, while preserving spatial hierarchy. The input frame $\mathrm{x} \in \mathrm{R}^{H\times W\times C}$ pass to the encoder, which transforms it into a latent representation $\mathrm{z} \in \mathrm{R}^{d}$, through a series of transformations. This introduces stochasticity in the latent space using a Variational Autoencoder (VAE) framework, where it learns a mean $\mu \in \mathrm{R}^{d}$ and a standard deviation $\sigma \in \mathrm{R}^{d}$ from the encoded features are calculated using Eq. (1) [35]:

$$\mu = f_{\mu}(x), \quad \log \sigma^2 = f_{\sigma}(x) \tag{1}$$

A sample latent vector z using the reparameterization trick is as shown in Eq. (2) [35]:

$$\mathrm{z} = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2}$$

This also allows for gradient-based optimization as well as introducing stochasticity into the latent space. The decoder takes the latent vector and reconstructs the input frame $\hat{x} = g(\mathrm{z})$, trying to replicate the original frame using transposed convolutional layers. CVAE (Convolutional Variational Autoencoder) minimizes a combined loss which contains two components: Reconstruction loss that measures how well the model can reproduce the original input using Eq. (3) [42]:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N} |\, x_i - \hat{x}_i|^2 \times N = \sum_{i=1}^{N} |\, x_i - \hat{x}_i|^2 \tag{3}$$

Kullback-Leibler Divergence calculated using Eq. (4) [35] Regularizes the latent space by encouraging the approximate posterior to match the standard normal distribution:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\sum_{i=1}^{d}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2) \tag{4}$$

The total loss encourages the model to accurately reconstruct input images while enforcing a regularized, continuous latent space for improved generalization in Eq. (5) [43]:

$$\mathcal{L}_{\text{CVAE}} = E_{x\sim p_{\text{dt}}}[\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}}] \tag{5}$$

After training on normal traffic frames, the model is evaluated on test frames. For each frame x_{j} , the reconstruction error E_{x_j} is computed using the mean squared error as in Eq. (6) [35]:

$$E_{x_j} = \frac{1}{HWC}\sum_{h,w,c}\left(x_j(h,w,c) - \hat{x}_j(h,w,c)\right)^2 \tag{6}$$

To interpret reconstruction errors as anomaly likelihood, the errors are scaled into regularity scores $R(x_j)\in[0,1]$, where higher scores indicate stronger similarity to normal data and lower scores suggest anomalies. The regularity score is computed as in Eq. (7) [35,36]:

$$R(x_j) = 1 - \frac{E_{x_j} - \min(E)}{\max(E) - \min(E) + \epsilon} \tag{7}$$

Here, $\min(E)$ and $\max(E)$ are the minimum and maximum reconstruction errors computed over the entire test set. ϵ is a small positive constant added for

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

numerical stability. Based on these scores, a binary classification is performed using a threshold τ to distinguish normal from anomalous events as in Eq. (8) [37]:

$$Anomaly(x_j) = \begin{cases} 1, & if\ R(x_j) < \tau \\ 0, & otherwise \end{cases} \quad (8)$$

To determine an appropriate threshold τ in Eq. (9) [42], we adopt a statistical thresholding approach based on the distribution of reconstruction errors on normal data:

$$\tau = \mu_{err} + 2 \cdot \sigma_{err} \quad (9)$$

Here, μ_{err} and σ_{err} are the mean and standard deviation of reconstruction errors from the normal training set. This approach assumes that normal instances are well reconstructed, while anomalies exhibit higher reconstruction errors and consequently lower regularity scores. The selected threshold τ directly influences the trade-off between false positives and false negatives. A lower threshold value may classify more frames as anomalous, increasing recall but potentially reducing precision due to higher false positive rates. Conversely, an increased threshold will lead to more conservative detection, which will increase precision and may overlook subtle accident events, which will reduce recall. The $2 \cdot \sigma_{err}$ value was selected empirically in this publication to provide a balanced trade-off between accuracy and recall, as shown in the published evaluation measures. To further prove the choice of the best threshold, a sensitivity analysis was made by changing the scaling factor of the standard deviation of the reconstruction errors. Thresholds in the range of $\mu_{err} + 1.5 \cdot \sigma_{err}$ to $\mu_{err} + 3 \cdot \sigma_{err}$ were evaluated on the validation set. Lower values favored higher recalls at the cost of increased false positives, while higher values improved precision but missed subtle anomalies. The threshold of $\mu_{err} + 2 \cdot \sigma_{err}$ consistently provided the best balance, achieving the highest F1-score and a stable trade-off between precision and recall. Therefore, this value was selected for all experimental evaluations.

F. Latent Space Clustering for Summarization

Once the frames of interest regarding an accident have been identified through the anomaly detection pipeline, a process of summarization goes into effect to create a concise yet informative summary of the video. This is done through the exploitation of the latent embeddings retrieved by the Conv-VAE that captures high-level structural features when projected to a reduced-dimensional space. Let the sequence of input video frames be denoted as $\{x_1, x_2, ..., x_t\}$ where each frame $x_t \in \mathrm{R}^{H \times W \times C}$ represents the video frame at time t. These frames are mapped to latent space using the encoder function f_{enc} as in Eq. (10) [37]:

$$z_t = f_{enc}(x_t), \text{ for } t = 1, ..., T \quad (10)$$

Here, $z_t \in \mathrm{R}^d$ denotes the d-dimensional latent embedding for frame x_t. To capture the diversity of visual content, the latent representations $\{z_1, z_2, \ldots, z_T\}$ are clustered using the K-Means. The objective of clustering is to partition the latent space into clusters $\{c_1, c_2, \ldots, c_k\}$ by minimizing the intra-cluster variance as follows in Eq. (11) [37]:

$$\min_{\{\mu_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{z \in C_k} ||z - \mu_k||^2 \quad (11)$$

where μ_k is the centroid of the cluster C_k. The number of clusters K determine the level of summarization and is chosen empirically based on the desired granularity. From each cluster C_k, a representative keyframe is selected by identifying the latent point $z_k^* \in C_k$ in Eq. (12) [35,37] that lies closest to the cluster centroid μ_k:

$$z_k^* = \arg\min_{z \in C_k} ||z - \mu_k||^2 \quad (12)$$

The corresponding input frame x_k^* is then designated as the keyframe for cluster C_k. This ensures that the selected keyframes are structurally representative of their respective visual contexts within the video. In this work, the number of clusters K is not fixed arbitrarily, but it is selected empirically based on the distribution and diversity of the accident-related frames in each video. Since the primary objective is video summarization rather than fine-grained classification, a relatively small and compact value of K is preferred in order to avoid unnecessary fragmentation of similar events. To ensure that the selected value of K produces meaningful and well-separated clusters, and silhouette analysis is used as a cluster validation technique. The silhouette score measures the similarity of a frame's latent representation to its own cluster compared to other clusters. Higher silhouette values indicate better cluster compactness and separation. The value of K that maximizes the average silhouette score is chosen as the optimal number of clusters for that video sequence. This adaptive selection strategy ensures that the resulting clusters exhibit strong compactness and separation, leading to representative and non-redundant keyframes in the final summary. To eliminate redundant frames that may carry visually similar content, pairwise structural similarity is computed between candidate keyframes using the Structural Similarity Index (SSIM) in Eq. (13) [35,37]:

$$\mathrm{SSIM}(x_i, x_j) = \frac{(2\mu_i\mu_j + C_1)(2\sigma_{ij} + C_2)}{(\mu_i^2 + \mu_j^2 + C_1)(\sigma_i^2 + \sigma_j^2 + C_2)} \quad (13)$$

where μ_i, σ_i^2, and σ_{ij} denote the mean, variance, and covariance of pixel intensities in grayscale images x_i and x_j, respectively. As a way of limiting repetition in time and to enhance processing power, frames whose Structural Similarity Index (SSIM) is greater than a predetermined value are deemed visually redundant

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

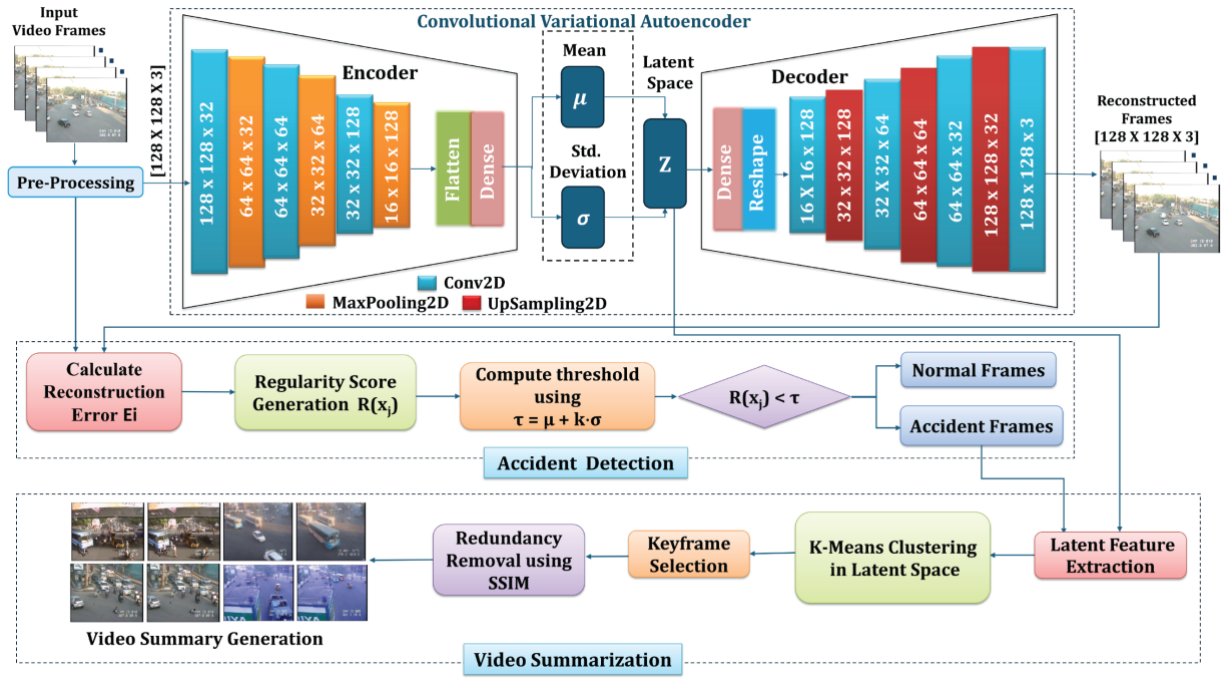


Fig. 3. Detailed architecture of the Convolutional Variational Autoencoder used for accident detection and video summarization.

and are not processed any further. The last summary video was assembled by arranging the chosen keyframes in chronological sequence and assembling them with a frame rate of 5 frames per second to facilitate qualitative analysis.

G. Dataset Discussion

The dataset to be applied in this research is the IITH road accident dataset, which was gathered through the CCTV surveillance system in Hyderabad City, India. In both video clips, the timing starts a few minutes prior to the event that an accident takes place and extends several minutes after [12]. There was a total of 94, 720 normal frames used in training, with 33,280 frames used in testing, consisting of 32,417 normal frames and 863 accident frames. Fig. 4 gifts displays sample records from the IITH data. UCF-Crime dataset is a massive video anomaly detection benchmark with 13 categories of real world incidents. In this case, we narrowed our research to the road accident subset, consisting of 150 CCTV videos obtained at varying times of day and night conditions, and with varying background scenes. Out of these, 127 were used for training and 23 for testing [48]. The dataset also has a high number of normal videos, and they can be well trained in one-class anomaly detection models.

H. Evaluation Metrics

In this section, we provide the evaluation metrics for our proposed Conv-VAE-based framework, both in the accuracy of the accident detection task and the keyframe summarization method. To determine the performance of Accident detection and latent clustering-based analysis of summarization, we used quantitative measures. The evaluation metric was chosen to capture both technical and practical performance of the surveillance systems in traffic. Although accuracy is limited to overall correctness, it is insufficient when there is an imbalance, as in accident detection. Accuracy is thus applied in order to decrease false alarms and recall to guarantee that real accidents are not overlooked. The F1-score offers a moderate perspective of both. The AUC shows that the model is capable of making a distinction between normal and accident frames according to the distribution of reconstruction scores, whereas the EER demonstrates the compromise between false positives and false negatives.

To be able to summarize, PSNR measures the quality of reconstruction, making sure that the normal

Fig. 4. Sample video frames from IITH Dataset.

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

patterns are learned correctly and that the accident frames result in more reconstruction errors. The rate of reduction indicates the level at which the video is condensed, and it is significant with regard to time and storage. The diversity rate (1 -SSIM) is used to guarantee that the keyframes used are not duplicates in terms of visual appearance. The silhouette score is used to describe the quality of clustering, which is based on the ability of the frames to cluster well in the latent space. The percentage of coverage ensures that the significant portions of the chain of accidents are not overlooked. Combined, these measurements indicate that the produced summaries are small, informative, and applicable to the real world, such as traffic monitoring, incident review, and emergency analysis.

1. Accident Detection Metrics

In an effort to assess the model in terms of distinguishing the accident and the normal frames, we utilized standard classification measures. Accuracy indicates the general percentage of correctly identified frames. Precision is the measure of the number of predicted frames which are really accidents and recall the number of actual accidents which have been identified [11]. F1-score is a harmonic mean of precision and recall, which offers a balanced index when the false positives or false negatives are both of interest [11]. We would also take the Equal Error Rate (EER) that is the error rate at which the false positive rate is the same as the false negative rate and this provides information on the trade-off between the two errors that the model can make. The curve of the ROC indicates the level at which the model distinguishes between accident and normal frames by displaying the true positive rate and false positive rate at different thresholds. This performance is summarized in the AUC score; the higher the value, the closer it approaches 1, the higher the performance in differentiating the two classes.

2. Latent Space Clustering-Based Summarization Evaluation

In order to extract important events in the sequence of accidents, we use clustering on the latent space of the trained Conv-VAE. The evaluation will be a combination of quantitative and qualitative analysis to make the summaries small, varied and representative.

A formal user study was not carried out in this work, but the chosen metrics of summarization are aimed at their close reflection in human perception of a good summary. Quantitatively, the rate of reduction is used to evaluate the compactness, and the diversity rate (1 − SSIM) is used to evaluate visual diversity among selected frames. The silhouette score, which is computed on original latent representations, is used to assess the quality of clustering, as well as aid in the decision of an acceptable number of clusters. The percentage coverage guarantees that the chosen frames will cover the whole temporal range of the incident. Also, PSNR is used to determine the visual fidelity of reconstructed frames by evaluating the pixel-by-pixel similarity of the reconstructed images to the original images. On the qualitative level, we make use of t-SNE-based latent space visualization to visualize the segregation and classification of accident and normal frames intuitively, which justifies the utility of our summarization strategy. A combination of these measures would give an approximation of what the human would consider a summary, as far as completeness, conciseness, and informativeness. The good qualitative correspondence between these quantitative findings and the visual summaries also justifies the usefulness of the proposed approach that can be applied in practice.

IV. Results

This section presents the outcomes regarding our accident detection and video summarization model, built using a latent space clustering approach with a Conv-VAE model.

A. Accident Detection Performance

In this subsection, we will emphasize the results of our model of accident detection using the Conv-VAE architecture on the IITH Accident Dataset and the UCF-Crime road-accident subset. Even though we train our model in an unsupervised fashion and only use normal frames in the training process, we manually annotated video frames of the IITH dataset with the characteristics of a normal or an accident to facilitate a quantitative analysis. We therefore report on key performance indicators, such as F1-score, accuracy, precision, recall, AUC, and Equal Error Rate (EER). Conversely, the UCF-Crime dataset contains the temporal annotations, which are predefined and show that there are some anomalous events in each video segment, which can be road accidents. This dataset therefore does not require manual labelling. These annotations allow one to directly and consistently assess the detection abilities of the model on unconstrained video data of the real world. Table 2 shows some important performance measures of the model, such as Accuracy, Precision, Recall, F1-Score, AUC, and EER. Fig. 5 provides a visual comparison of these performance metrics across both datasets, clearly illustrating the effectiveness of our approach across different data sources. Additionally, the Receiver Operating Characteristic (ROC) curves for both (a) the IITH dataset and (b) the UCF-Crime dataset are shown in Fig. 6, offering further insight into the model's discriminative capabilities. A short examination of the erroneously classified frames reveals that false positives are mostly found in sudden but not accidental

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Table 2. Performance comparison of the proposed method on the IIT Hyderabad and UCF-Crime datasets.

Metric	IIT Hyderabad Dataset	UCF-Crime Dataset
Accuracy	93.5%	91.2%
Precision	0.874%	0.845%
Recall	0.801%	0.768%
F1-Score	0.836%	0.804%
AUC	0.9061	0.8795
EER	18.2%	21.4%

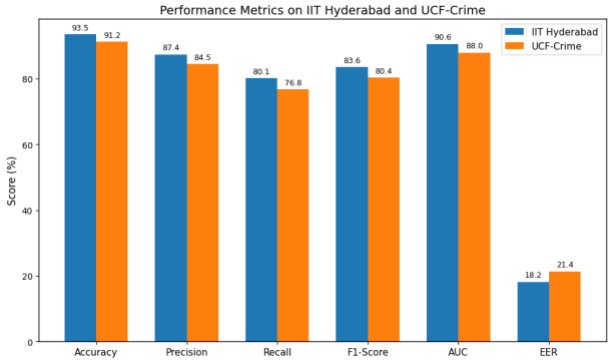


Fig. 5. Sample video frames from IITH Dataset.

visual events such as sudden braking,sharp lane turns, and heavy occlusions, which enhance the error in reconstruction and reduce accuracy (0.874 and 0.845 in IITH and UCF-Crime, respectively). False negatives, on the other hand, are largely due to small or far collisions and low visibility (e.g., low light or motion blur), which yields low recall values of 0.801 (IITH) and 0.768 (UCF-Crime).

These results suggest that CVAE-ADS is useful in identifying visually salient accidents but might fail to pick up finer details, which implies that the use of a temporal model and attention could help to enhance performance. To assess the stability of the proposed CVAE-ADS framework, all the experiments were run repeatedly with various random initializations as demonstrated in Table 3. The results reported on the performance metrics are the average values throughout these runs, as well as standard deviations. The differences in AUC and F1-score were also not very significant (within +1-2 percent), and this implies that the model gives consistent and reliable results every time it is run. This proves that the reported performance is not an outcome of random chance, and it shows the strength of the suggested course of action. Since these metrics capture overall model behavior,

Table 3. Performance Stability of CVAE-ADS Across Multiple Runs

Metric	IITH (Mean ± Std)	UCF-Crime (Mean ± Std)
Accuracy	93.5 ± 0.8	91.2 ± 1.1
AUC	90.61 ± 1.02	87.95 ± 1.28
F1-score	0.836 ± 0.03	0.804 ± 0.04

variability is reported only for Accuracy, AUC, and F1-score, while the remaining metrics showed consistent trends across runs.

B. Threshold Sensitivity Analysis

To further examine the influence of the reconstruction error threshold on detection performance, a sensitivity analysis was conducted by varying the threshold from

Table 4. Comparison of model performance metrics across different statistical threshold levels.

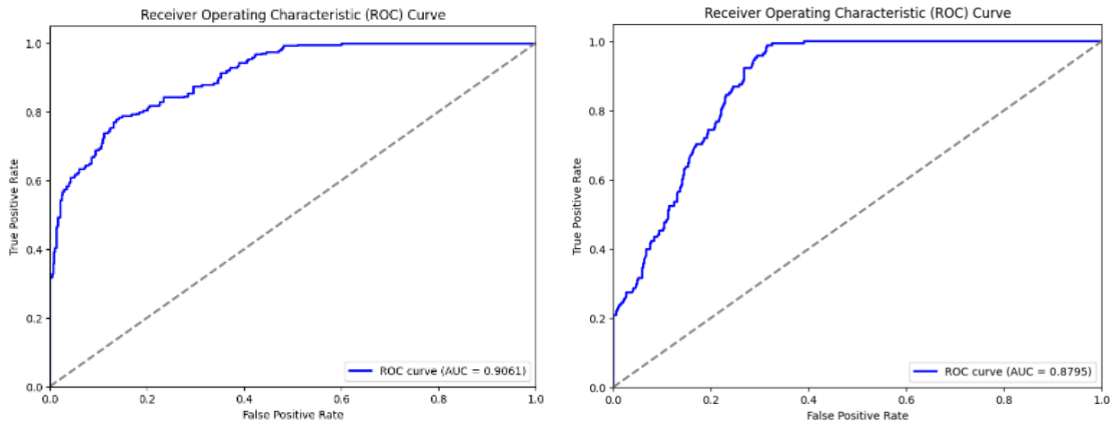
k value	Threshold	Precision	Recall	F1-score
1.5	μ + 1.5σ	0.74	0.86	0.79
2.0	μ + 2σ	0.87	0.80	0.83
2.5	μ + 2.5σ	0.91	0.71	0.80
3.0	μ + 3σ	0.95	0.62	0.75

$\mu_{err} + 1.5 \cdot \sigma_{err}$ to $\mu_{err} + 3 \cdot \sigma_{err}$. The corresponding changes in Precision, Recall, and F1-score are reported in Table 4. As shown, the threshold of $\mu_{err} + 2 \cdot \sigma_{err}$ achieves the most balanced trade-off between false positives and false negatives and is therefore selected for all final evaluations.

C. Latent Space-based Video Summarization Performance

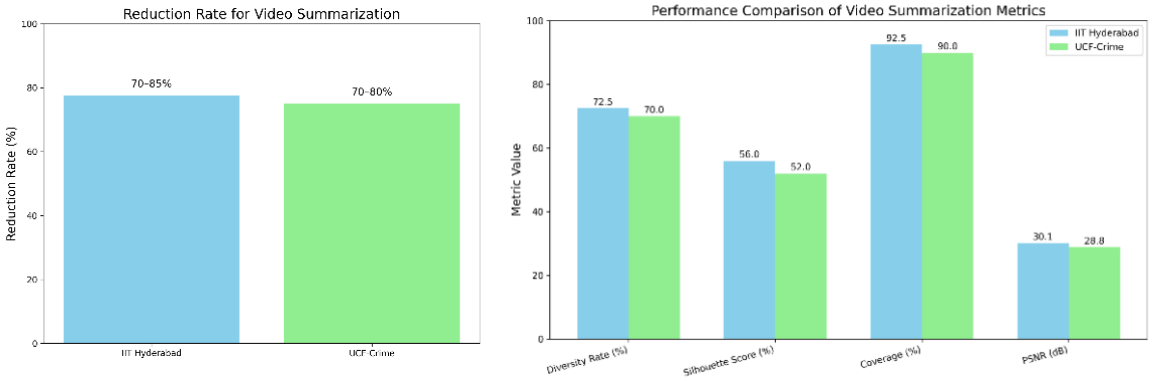
This section is a report of the findings of our keyframe summarization process, as is presented in Table 5. The latent space clustering performance is demonstrated by the performance of the latent space clustering on summarizing diverse and concise summaries. The approach was evaluated on the IITH and UCF-Crime datasets to confirm its effectiveness. Fig. 7 summarizes the relative results of our summarization measures on both datasets and demonstrates that the proposed diversity, and coverage. The performance of video summarization usually differs depending on the sequences based on the duration of the sequences, movement, and events. CVAE-ADS model preserves high reduction rates, Our reduction rate was 7085% based on video length and content complexity, with coverage (greater than 90) and perceptual quality (PSNR = 28.8-30.1 dB). Furthermore, Fig. 8 provides a visual comparison between the original and

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139



(a) (b)

Fig. 6. ROC curves for the proposed accident detection model on (a) the IITH Dataset and (b) the UCF-Crime Dataset.



(a) (b)

Fig. 7. Comparison of video summarization results on IIT Hyderabad and UCF-Crime datasets. (a) Reduction Rate (%) comparison on IITH and UCF-Crime datasets. (b) Other summarization metrics showing diversity, clustering, coverage, and quality performance.

reconstructed frames. For normal traffic scenes, the model produces reconstructions that are visually close to the original frames, consistent with the high PSNR values. In contrast, accident and abnormal frames show noticeable distortion and blurring in the reconstructed output, resulting in higher reconstruction error. This behavior validates the effectiveness of the reconstruction-based anomaly detection strategy, where abnormal events naturally lead to higher error and lower regularity scores. Together, these visual results provide strong qualitative evidence that supports the quantitative performance metrics and confirms the model's ability to both detect anomalies and generate meaningful summaries under real-world traffic conditions. Fig. 9(a) and 9(b) present the t-SNE visualization of latent features extracted by the proposed CVAE-ADS for the IITH and UCF-Crime datasets, respectively. In both cases, normal and accident frames form visibly distinct clusters, indicating that the model has learned a well-structured latent space. This clear separation directly supports the high

Table 5. Comparison of video summarization metrics between the IITH and UCF-Crime datasets.

Metric	IIT Hyderabad Dataset	UCF-Crime Dataset
Reduction Rate (%)	70-85%	70–80%
Diversity Rate (1 − SSIM)	0.7249	0.70
Silhouette Score (Latent Space)	0.56	0.52
Coverage Percentage (%)	92.5%	90.0%
PSNR	30.1 dB	28.8 dB

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Table 6. Comparison of state-of-the-art accident detection methods with the proposed CVAE-ADS Model.

Author(s)	Learning	Approach	Dataset	AUC (%)	EER (%)	mAP (%)	Detecti on Rate	FAR
D. Singh et al. [12]	Unsupervised	Unsupervised Denoising Autoencoder + One-Class SVM	IITH	77	22.50	-	-	-
Srinivasan et al. [13]	Supervised	DETR + Random Forest		82	-	-	78.2%	-
Wang et al. [14]	Supervised	Retinex + YOLOv3 + Decision Tree	Online CCTV videos	96.32	-	-	92.5%	7.5
Khan et al. [16]	Supervised	CNN + Rolling Prediction	VAID & Test Dataset	-	-	-	88%	-
Pawar et al. [17]	Unsupervised	Conv-AE + Seq2Seq LSTM Autoencoder	IITH	79	20.50	60	-	-
			DOTA	84.70	11.7	-	-	-
A. R. Pathak et al. [18]	Supervised	YOLOv2 + Transfer Learning	IITH	-	-	76	-	-
Thakare et al. [19]	Semi-Supervised	Object Interaction + Refinement + Heatmaps	UCF Crime	69.70	-	-	-	0.8
			CADP	72.59	-	-	-	2.2
Adewopo et al. [20]	Supervised	Lightweight I3D - ConvLSTM2D	Custom Dataset	-	-	87	80%	-
Chauhan A et al.	Unsupervised	Proposed CVAE-ADS	IITH	90.61	18.2	-	-	-
			UCF-Crime	87.95	21.4	-	-	-

AUC values (90.61% for IITH and 87.95% for UCF-Crime) and the strong silhouette scores reported in Table 5, confirming that the learned representations are both compact and discriminative. At the same time, some overlap between clusters can still be observed in complex lighting and heavy occlusion scenarios, indicating a potential area for improvement through explicit temporal modeling in future work. Additionally, the output summaries generated from the selected keyframes are illustrated in Fig. 10 (a) and 10 (b) for both the IITH and UCF-Crime datasets, showcasing the model's ability to produce concise and representative video summaries.

D. Comparative Analysis with State-of-the-Art

We evaluated our proposed CVE-ADS model against existing methods for accident detection and video summarization. Table 6 compares our offered CVAE-ADS model with the current methods of accident detection on various datasets. The model had a significantly greater AUC of 90.61% on IITH and 87.95% on UCF-Crime, and was better than prior choices of an unsupervised method. CVAE-ADS is also more effective in identifying rare accident cases without the use of labelled anomaly information due to the combination of convolutional feature representation and variational learning. Though Wang et al. [14] mention a high AUC of 96.32% with a supervised method that uses Retinex, YOLOv3, and a decision tree, their approach is based on the use of labelled data on accidents and the need to optimize lighting conditions to fit CCTV cameras. Conversely, CVAE-ADS operates in an unsupervised fashion and does not need annotated data to be available a priori, which makes it more scalable in real-world applications. Compared to the other unsupervised methods, such as Singh et al. [12] with 77% and Pawar et al. [17] with 79% AUC on the IITH dataset, CVAE-ADS is much better with a 90.61% AUC. This advancement shows its capability to study the dynamics of scenes involved in the complex scene and other

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Table 7. Comparison of proposed CVAF-ADS With state-of-the-art video summarization approaches

Author(s)	Approach / Model	Dataset(s) Used	Reduction Rate (%)	Diversity Rate	Coverage (%)	PSNR (dB)
Thomas et al. [7]	Perceptual Video Summarization	YouTube-8M, Urban Tracker	Used saliency	×	×	×
Kosambia et al. [21]	Video Synopsis using ResNet	IITH	×	×	×	×
Pramanik et al. [22]	Z-STRFG: Spatio-temporal fuzzy approach	YouTube8M, Anomaly20, Urban Tracker	×	×	×	×
Mehwish Tahir et al. [23]	YOLOv5 + Privacy-preserved summarization	Synthetic + Real-time Data	42.97%	×	×	×
Saxena et al. [24]	YOLOv5 + Event-based summarization	Synthetic + Real Traffic Videos	20–50%	×	×	×
Proposed CVAE-ADS Approach	Conv-VAE + Latent Space Clustering	IITH,	70-85%	0.7249	92.5%	30.1
		UCF-Crime	70-80%	0.70	90.0%	28.8

minute abnormalities that may be difficult to detect using the manual.

The superior performance of the proposed CVAE-ADS is largely attributed to its ability to learn a compact and structured representation of normal traffic patterns without relying on labeled accident data. Unlike conventional autoencoders or CNN-based classifiers, the variational formulation enforces regularization in the latent space, enabling clearer separation between normal and anomalous patterns. In addition, the convolutional layers preserve spatial characteristics such as vehicle structure and collision regions, which are often lost in fully connected or handcrafted feature-based methods. This architectural combination of probabilistic modeling and spatial feature learning enhances class separability, resulting in higher AUC and lower EER on both the IITH and UCF-Crime datasets, making the proposed approach more robust and scalable for real-world traffic surveillance applications. In terms of video summarization, our method extracts keyframes that cover not only scene variety but also critical accident moments. Compared to most of the prior works concentrating on random or uniform sampling, our approach can select frames not only as representatives of scene diversity but also enclosing critical accident moments. Here, we compared its performance using the key evaluation metrics with existing state-of-the-art summarization techniques, as shown in Table 7. Experiments demonstrate that the coverage and compression efficiency of CVAE-ADS outperform existing summarization methods. While methods like YOLO-based summarization [23][24] achieve intermediate reduction rates, they do not present detailed metrics, e.g., diversity or coverage. On the other hand, CVAE-ADS can achieve up to 85% reduction rates with 92.5% of coverage, and we obtain strong diversity scores as well that outperform previous works such as Thomas et al. [7] and Pramanik et al. [22] that do not provide

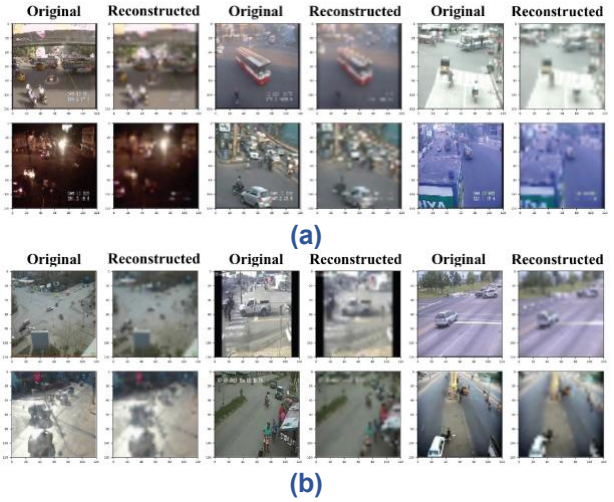


Fig. 8. Visual comparison between original and reconstructed frames generated by the proposed CVAE-ADS model. (a) Results on the IITH Dataset. (b) Results on the UCF-Crime Dataset.

comprehensive quantitative results. This demonstrates the proposed approach for how it effectively produces concise and informative summaries. The effectiveness of the CVAE-ADS is also shown through the comparison with existing approaches since it delivers a good overall result in accident detection and video summarization. Our model improves reduction rates while demonstrating better coverage with diverse keyframe selection without the need for labelled data. The visualization of the clearly separate latent space demonstrates that the model represents meaningful scene dynamics successfully. Much existing research

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

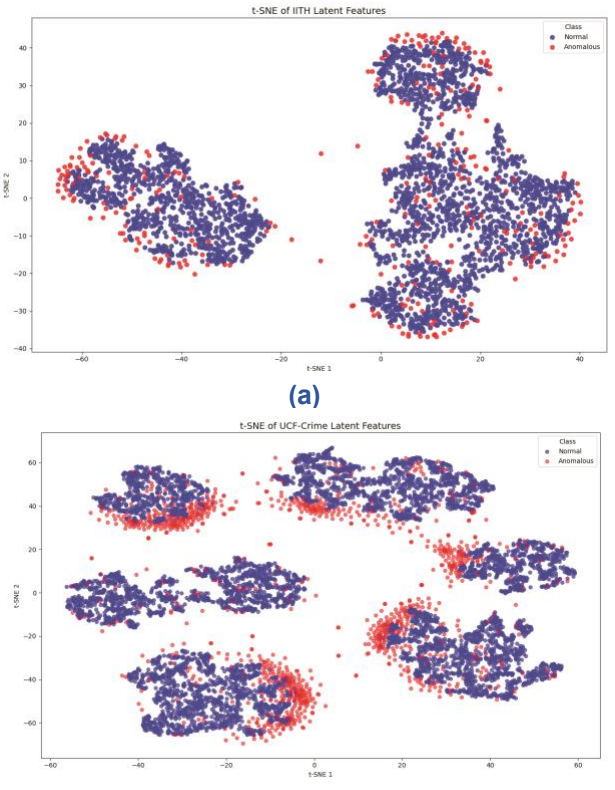


Fig. 9. Latent Space Clustering via t-SNE for (a) IITH Dataset and (b) UCF-crime Dataset

relies on synthetic datasets or fails to evaluate crucial performance metrics. The proposed CVAE-ADS model, evaluated on real-world datasets, proves to be a scalable as well as practical solution for intelligent traffic surveillance systems. Despite CVAE-ADS being tested in an offline experimental environment, its compact convolutional architecture and frame-by-frame analysis can be used in real-time in the conditions of real traffic surveillance. No complicated object tracking or external detectors are necessary with the model, reducing the computational overhead and allowing faster inference with conventional GPU-enabled monitoring systems. This enables the accident occurrences to be identified in time, which is helpful for the production of quick alerts to the traffic control rooms and emergency response teams. Moreover, automatic summarization of keyframes reduces the workload of video inspectors, enabling authorities to assess the extent of an incident within seconds and the context of the situation to make decisions efficiently.

The existing CVAE-ADS framework mainly works with single video frames, and it aims at learning spatial representations of normal traffic scenes to detect anomalies. Although this method is good at detecting visually sudden and structural anomalies, it does not directly model time-based dependence or dynamic motion patterns of successive frames, which can be informative in complex accident situations that have slow transitions or dynamic effects. To overcome this weakness, the proposed framework will be further expanded in the future with the inclusion of the temporal modeling mechanisms, including Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), or 3D convolutional networks to learn both spatial and temporal correlations in video sequences. Such temporal components are predicted to improve the detection of the subtle and time-varying anomalies as well as the increased robustness of the system in real-world traffic conditions. In addition, the experiments presented in this study are limited to the IITH and UCF-Crime datasets, which mainly consist of fixed-angle CCTV footage captured in urban and highway environments. These datasets do not fully represent all real-world conditions, such as rural roads, extreme weather, varying camera heights, wide-angle lenses, or dense heterogeneous traffic patterns. As a result, the model may encounter challenges when deployed in unseen environments with significant visual or contextual differences. Although the unsupervised design improves adaptability, further evaluation on more diverse and geographically distributed datasets is necessary to strengthen the model's generalization capability.

V. Discussion

The experiments demonstrate that CVAE-ADS provides a competitive and robust solution for both accident detection and summarization in the context of traffic surveillance systems. The reconstruction-based anomaly scoring mechanism is able to recognize abnormal frames with high accuracy, while the latent space clustering approach retains the most informative segments for efficient summarization. The following discussion discusses how these findings relate to existing literature, pointing out similarities and contradictions, along with presenting current limitations of the proposed system. Compared with previous deep autoencoder-based anomaly detection frameworks such as those reported in [12], CVAE-ADS achieves a higher AUC, with an improvement of 13.6% on average. This is mainly due to its probabilistic latent modeling capability. Unlike stacked denoising autoencoders learning deterministic embeddings, our conditional latent space provides better separation between normal and accident events. Unlike approaches based on supervised learning or extensive labeled datasets, as in [14], CVAE-ADS does not need any manual annotations or pre-processing operations like Retinex enhancement. Hence, its scalability and applicability to realistic traffic surveillance environments significantly improve. Beyond [18], the proposed approach outperforms a temporal CNN-based architecture, which easily results in overfitting

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

when trained with limited accident datasets. Moreover, the reconstruction-driven anomaly metric of our framework generalizes better to unseen scenarios.

Our results are consistent with findings in [22], which emphasized that the latent-space structure is critical for enhanced anomaly sensitivity in video surveillance applications. As in their studies, we found that latent-space clustering captures contextual representations relevant to summarizing critical frames.

However, our results run counter to detection-only approaches using YOLO-based event localization networks as in [23] and [24]. These supervised detection systems tend to suffer from low-light conditions, partial occlusion, and motion blur, commonly occurring in traffic-camera videos. By contrast, CVAE-ADS shows consistent detection performance as the reconstruction error is inherently less sensitive to illumination variations and does not depend on explicit bounding-box predictions. Likewise,

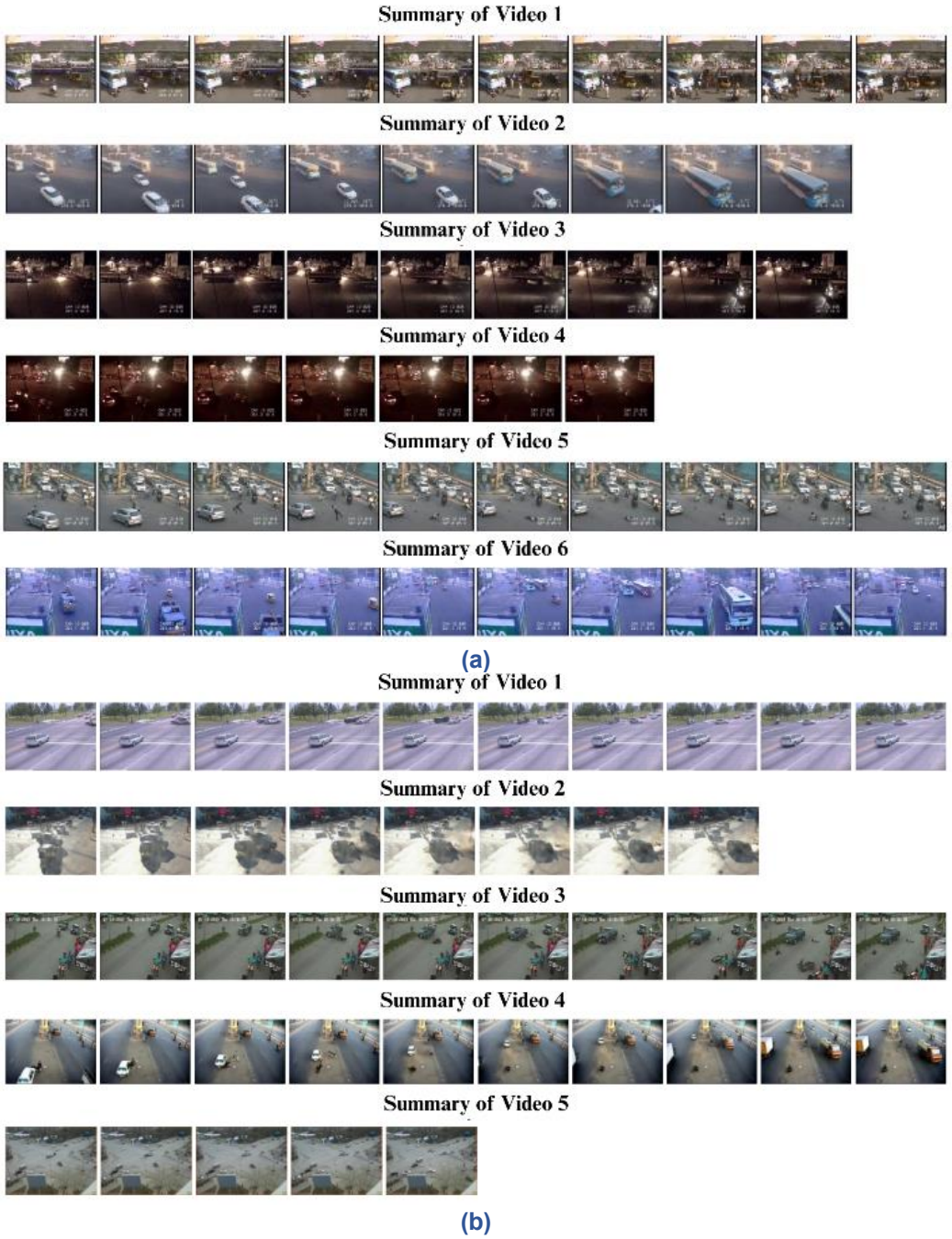


Fig. 10. Sample video summarization results generated by the proposed CVAE-ADS model. (a) IITH accident dataset (b) UCF-Crime Dataset

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

handcrafted-feature-based or purely motion-heuristic-based methods, as in [26], fail to generalize well to different accident types; our latent-space learning approach captures subtle deviations of spatial–appearance cues much better.

Despite its effectiveness, the CVAE-ADS framework has certain limitations. The current architecture, for starters, does not make use of explicit temporal modeling mechanisms such as LSTM networks, 3D CNNs, and transformer-based architectures. Due to this, the sequential motion dependencies between frames are not entirely captured, which may result in failing to detect accidents that evolve slowly or are subtle in nature. Moreover, the model is sensitive to challenging environmental conditions such as very low illumination, severe occlusions, and camera shake that badly affect the discriminative power of the reconstruction-based score for anomaly detection. Additionally, the experimental assessment is restricted to two benchmark datasets, namely the IITH and UCF-Crime datasets, which contain fixed-angle CCTV surveillance videos. As a result, the model has not undergone validation on a broad spectrum of scenarios, including drone-based surveillance and multi-camera systems or wide-area traffic monitoring environments. While the computational efficiency of CVAE-ADS is promising, we have yet to assess its real-time deployment on embedded edge devices, resource-constrained GPGPU platforms, or large-scale smart cities. In conclusion, the existing framework does not include mechanisms for dynamic adaptation to scenes and domain transfer, which may lead to limitations in its scalability and generalization when used in various geographic, camera and traffic settings.

VI. Conclusion

In this work, we present CVAE-ADS, a lightweight, unsupervised Conditional Variational Autoencoder–based framework designed for effective accident detection and video summarization in traffic surveillance systems. The model consists of a two-stage pipeline: i) reconstruction-based anomaly scoring for accurate accident detection, and ii) latent-space clustering for generating compact and informative summaries that capture critical event information. Experimental results on both the IITH Accident Dataset and the Accident subset of UCF-Crime demonstrate the superior performance of CVAE-ADS against state-of-the-art methods. It achieved 93.5% accuracy and 90.61% AUC on the IITH dataset and 87.95% AUC on the UCF-Crime dataset. Furthermore, for video summarization, CVAE-ADS was able to reduce the video length by 70–85% while maintaining 92.5% coverage of accident-related segments.

These results validate that CVAE-ADS can effectively detect anomalies and generate short summaries that greatly enhance the efficiency of post-incident review, traffic monitoring, emergency response workflows, and legal verification processes. In future work, we would like to extend CVAE-ADS by introducing temporal modeling using 3D-CNNs, LSTMs, or transformer-based architectures to capture more complex motion patterns and contextual cues related to accidents. We also want to try GAN-enhanced reconstruction, multi-camera fusion, and deployment in large-scale real-world scenarios across diverse and challenging traffic conditions to further validate its generalization capability.

Acknowledgment

We sincerely thank CVM University and Gujarat Technological University for providing the essential infrastructure, academic support, and research facilities that made this work possible. We also acknowledge the contributors and maintainers of publicly available traffic accident datasets, whose resources were invaluable for training and evaluating our CVAE-ADS framework. Our heartfelt appreciation goes to our faculty mentor, research colleagues, and technical staff for their continuous encouragement, insightful suggestions, and assistance throughout the course of this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets used in this study are publicly available benchmark datasets and can be accessed from the following sources:

- IIT Hyderabad Road Accident Dataset (IITH): This dataset consists of real-world traffic accident surveillance videos collected from CCTV cameras in Hyderabad, India. It is publicly available and was accessed from: https://sites.google.com/site/dineshsinghindian/iith_accident-dataset
- UCF-Crime Dataset (Road Accident Subset): The UCF-Crime dataset is a large-scale video anomaly detection benchmark containing real-world surveillance videos. In this work, only the road accident subset was used. The dataset is publicly available at: https://www.crcv.ucf.edu/research/real-world-anomaly-detection-in-surveillance-videos/ A downloadable version is available via:

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

https://www.dropbox.com/scl/fo/2aczdnx37hxvcfdo4rq4q/AOjRokSTaiKxXmgUyqdcl6k

Author Contribution

Ankita Chauhan: Conceptualization, methodology, implementation of the CVAE-ADS framework, data preprocessing, model training and evaluation, experiments, and manuscript drafting. Dr. Sudhir Vegad: Supervision, technical validation, critical manuscript revision, contribution to related work review, and guidance on hybrid model design. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work to ensure integrity and accuracy.

Declarations

Ethical Approval

This study does not involve human participants, animals, or personally identifiable information. All datasets used are publicly available and have been utilized strictly for academic research purposes, in compliance with their respective terms of use.

Consent for Publication Participants.

Consent for publication was given by all participants

Competing Interests

The authors declare no competing interests.

References

[1] WHO, Road traffic injuries, Retrived 13 December 2023, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] Ministry of Road Transport and Highways. (2022). Road accidents in India 2022, https://morth.nic.in/sites/default/files/RA_2022_30_Oct.pdf

[3] Ministry of Road Transport and Highways. (2024, July 24). Deaths due to road accidents in India. Press Information Bureau. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2036268

[4] Tiezzi, M., Melacci, S., Maggini, M., Frosini, A. (2018). Video Surveillance of Highway Traffic Events by Deep Learning Architectures. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science(), vol 11141. Springer, Cham. https://doi.org/10.1007/978-3-030-01424-7_57

[5] 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

[6] P. Kadam et al., "Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms," in IEEE Access, vol. 10, pp. 122762-122785, 2022, doi: 10.1109/ACCESS.2022.3223379.ss

[7] S. S. Thomas, S. Gupta, and V. K. Subramanian, “Event detection on roads using perceptual video summarization,” IEEE Trans. Intell. Transp. Syst., vol. 19, no. 9, pp. 2944–2954, Dec. 2017.

[8] Adewopo, Victor & Elsayed, Nelly & Elsayed, Zag & Ozer, M. & Abdelgawad, Ahmed & Bayoumi, Magdy. (2022). Review on Action Recognition for Accident Detection in Smart City Transportation Systems.

[9] Saini, P., Kumar, K., Kashid, S. et al. Video summarization using deep learning techniques: a detailed analysis and investigation. Artif Intell Rev 56, 12347–12385 (2023). https://doi.org/10.1007/s10462-023-10444-0

[10] Pawar, K., Attar, V. Deep learning approaches for video-based anomalous activity detection. World Wide Web 22, 571–601 (2019). https://doi.org/10.1007/s11280-018-0582-1

[11] Gandhi, V., Chaudhari, Y., Kumar, A. *et al.* Benchmarking Machine Learning Models for Obesity Classification with SHAP-Based Interpretability. *Int J Comput Intell Syst* (2025). https://doi.org/10.1007/s44196-025-01078-x

[12] D. Singh and C. K. Mohan, "Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 3, pp. 879-887, March 2019, doi: 10.1109/TITS.2018.2835308.

[13] 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.

[14] 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

[15] Robles-Serrano, Sergio, German Sanchez-Torres, and John Branch-Bedoya. 2021. "Automatic Detection of Traffic Accidents from Video Using Deep Learning Techniques" Computers 10, no. 11: 148. https://doi.org/10.3390/computers10110148

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

[16] 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

[17] 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.

[18] 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.

[19] 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

[20] 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.

[21] 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

[22] 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.

[23] 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

[24] 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

[25] Zhang, K., Chao, WL., Sha, F., Grauman, K. (2016). Video Summarization with Long Short-Term Memory. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9911. Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_47

[26] Mayya, V., Nayak, A.: Traffic surveillance video summarization for detecting traffic rules violators using R-CNN. In: Advances in Computer Communication and Computational Sciences—Proceedings of IC4S 2017, pp. 117–126. Springer Verlag (2019). https://doi.org/10.1007/978- 981-13-0341-8_11

[27] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P. (2019). Summarizing Videos with Attention. In: Carneiro, G., You, S. (eds) Computer Vision – ACCV 2018 Workshops. ACCV 2018. Lecture Notes in Computer Science(), vol 11367. Springer, Cham. https://doi.org/10.1007/978-3-030-21074-8_4

[28] Jingxu Lin, Sheng-hua Zhong, Ahmed Fares,Deep hierarchical LSTM networks with attention for video summarization, Comsputers & Electrical Engineering, Volume 97, 2022, 107618, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107618.

[29] Payal Kadam, Deepali Vora, Shruti Patil, Sashikala Mishra, Vaishali Khairnar, Behavioral profiling for adaptive video summarization: From generalization to personalization, MethodsX, Volume 13, 2024, 102780, ISSN 2215-0161, https://doi.org/10.1016/j.mex.2024.102780.

[30] Zhang, Yujia, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P. Xing. 2020. “Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder.” Pattern Recognition Letters 130: 376–85. https://doi.org/https://doi.org/10.1016/j.patrec.2018.07.030.

[31] Ji Zhong, Xiong K, Pang Y, Li X. Video summarization with attention-based encoder–decoder networks. IEEE Transactions on Circuits and Systems for Video Technology. 2019; 30(6):1709-17.

[32] Sheng-Hua Zhong, Jingxu Lin, Jianglin Lu, Ahmed Fares, and Tongwei Ren. 2022. Deep Semantic and Attentive Network for Unsupervised Video Summarization. ACM Trans. Multimedia Comput. Commun. Appl. 18, 2, Article 55 (May 2022), 21 pages. https://doi.org/10.1145/3477538

[33] Sreeja, M.U., Kovoor, B.C. A multi-stage deep adversarial network for video summarization with knowledge distillation. J Ambient Intell Human

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

Comput 14, 9823–9838 (2023). https://doi.org/10.1007/s12652-021-03641-8

[34] Y. Yuan and J. Zhang, "Unsupervised Video Summarization via Deep Reinforcement Learning With Shot-Level Semantics," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 1, pp. 445-456, Jan. 2023, doi: 10.1109/TCSVT.2022.3197819.

[35] Panda, Rameswar, Abir Das, Ziyan Wu, Jan Ernst and Amit K. Roy-Chowdhury. "Weakly Supervised Summarization of Web Videos." 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 3677-3686.

[36] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly-supervised video summarization using variational encoder–decoder and web prior," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 184–200.

[37] Ramos W, Silva M, Araujo E, Moura V, Oliveira K, Marcolino LS, Nascimento ER (2022) Text-driven video acceleration: a weakly-supervised reinforcement learning method. IEEE Trans Pattern Anal Mach Intell 45(2):2492–2504

[38] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 929, 7582–7589

[39] Xu J, Sun Z, Ma C (2021) Crowd aware summarization of surveillance videos by deep reinforcement learning. Multimed Tools Appl 80:6121–6141. https://doi.org/10.1007/s11042-020-09888-1

[40] Liu T, Meng Q, Huang J-J, Vlontzos A, Rueckert D, Kainz B (2022) Video summarization through reinforcement learning with a 3d spatio-temporal u-net. IEEE Trans Image Process 31:1573–1586

[41] Guolong Wang, Xun Wu, Junchi Yan, Progressive reinforcement learning for video summarization, Information Sciences, Volume 655, 2024, 119888, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2023.119888.

[42] An, Jinwon and Sungzoon Cho. "Variational Autoencoder based Anomaly Detection using Reconstruction Probability." (2015).

[43] Nguyen, H.H., Nguyen, C.N., Dao, X.T., Duong, Q.T., Kim, D.P., & Pham, M. (2024). Variational Autoencoder for Anomaly Detection: A Comparative Study. ArXiv, abs/2408.13561.

[44] Iqbal, T., & Qureshi, S. (2022). Reconstruction probability-based anomaly detection using variational auto-encoders. International Journal of Computers and Applications, 45(3), 231–237. https://doi.org/10.1080/1206212X.2022.2143026

[45] N. Aslam and M. H. Kolekar, "A-VAE: Attention based Variational Autoencoder for Traffic Video Anomaly Detection," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-7, doi: 10.1109/I2CT57861.2023.10126296.

[46] Wenhao Yu, Qinghong Huang, A deep encoder-decoder network for anomaly detection in driving trajectory behavior under spatio-temporal context, International Journal of Applied Earth Observation and Geoinformation, Volume 115, 2022, 103115, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2022.103115.

[47] Zhang, C., Wang, X., Zhang, J. et al. VESC: a new variational autoencoder based model for anomaly detection. Int. J. Mach. Learn. & Cyber. 14, 683–696 (2023). https://doi.org/10.1007/s13042-022-01657-ws

[48] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6479–6488

[49] Quang Nguyen Huy Minh, Nen Nguyen Dinh, Long Viet Ho, Cuong Phan Huu, Real-time traffic accident detection using yolov8, Transportation Research Procedia, Volume 85, 2025, Pages 68-75, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2025.03.135.

[50] Md Shamsul Arefin, Md Ibrahim Shikder Mahin, Farzana Akter Mily, Real-time rapid accident detection for optimizing road safety in Bangladesh, Heliyon, Volume 11, Issue 4, 2025, e42432, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2025.e42432.

Author Biography

Ankita Chauhan received her Master of Engineering degree in 2013 from Gujarat Technological University, India, and is currently pursuing her Doctor of Philosophy in Computer Engineering at the same university. She is an Assistant Professor at the Madhuben and Bhanubhai Patel Institute of Technology, CVM University, Gujarat, and has over 15 years of teaching experience. Her doctoral research focuses on artificial intelligence-driven video analytics, including deep learning based anomaly detection and traffic accident detection. Her research interests include machine learning, deep

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139

learning, computer vision, and intelligent surveillance systems. She has published research articles in reputed conferences and peer-reviewed journals.

Dr. Sudhir Vegad received his B.E. degree in Information Technology from Saurashtra University in 2002, his M.E. degree in Computer Engineering from Sardar Patel University, Vallabh Vidyanagar, in 2007, and his Ph.D. degree from Gujarat Technological University, Ahmedabad, in 2018. He is currently a Professor and Head of the Department of Computer Engineering at G. H. Patel College of Engineering & Technology, The Charutar Vidya Mandal (CVM) University, India. His research interests include Machine Learning, Deep Learning, Computer Vision, and Image Processing. He has published several research articles in peer-reviewed journals and reputable international conference proceedings and has delivered expert lectures in the domains of Image Processing and Computer Vision.

Manuscript received 16 August 2025; Revised 10 November 2025; Accepted 20 December 2025; Available online 1 January 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i1.1139