EVENT BASED VIDEO SUMMARIZATION FOR TRAFFIC SURVEILLANCE

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer / IT Engineering

by

Chauhan Ankita P.

[Enrollment No.: 199999913508]

under supervision of

Dr. Sudhir P Vegad

GUJARAT TECHNOLOGICAL UNIVERSITY

AHMEDABAD

[July – 2026]

EVENT BASED VIDEO SUMMARIZATION FOR TRAFFIC SURVEILLANCE

A Thesis submitted to Gujarat Technological University

for the Award of

Doctor of Philosophy

in

Computer / IT Engineering

by

Chauhan Ankita P.

[Enrollment No.: 199999913508]

under supervision of

Dr. Sudhir P Vegad

GUJARAT TECHNOLOGICAL UNIVERSITY

AHMEDABAD

[July – 2026]

DECLARATION

I declare that the thesis entitled **"Event based Video Summarization for Traffic Surveillance"** submitted by me for the degree of Doctor of Philosophy is the record of research work carried out by me during the period from December 2019 to May 2026 under the supervision of **Dr. Sudhir P Vegad** and this has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities, if noticed in the thesis.

Signature of the Research Scholar:  Date: 30th July 2026

Name of Research Scholar: **Chauhan Ankita P.**

Place: Anand

CERTIFICATE

I certify that the work incorporated in the thesis **"Event based Video Summarization for Traffic Surveillance"** submitted by **Chauhan Ankita P.** was carried out by the candidate under my supervision/guidance. To the best of my knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of the Research Supervisor:  Date: 30th July 2026

Name of Supervisor: Dr. Sudhir P Vegad

Place: Anand

COURSE-WORK COMPLETION CERTIFICATE

This is to certify that **Chauhan Ankita P.** enrolment no. **199999913508** is enrolled for PhD program in the branch **Computer/ IT Engineering** of Gujarat Technological University, Ahmedabad.

(Please tick the relevant option(s))

☐ ~~He/She has been exempted from the course-work (successfully completed during M.Phil Course)~~

☐ ~~He/She has been exempted from Research Methodology Course only (successfully completed during M.Phil Course)~~

☑ ~~He~~/She has successfully completed the PhD course work for the partial requirement for the award of PhD Degree. ~~His~~/ Her performance in the course work is as follows-

Grade Obtained in Research Methodology [PHD2001]	Grade Obtained in Research and Publication Ethics [PHD2002]	Grade Obtained in Self-Study Course [PHD2003]
AB	AA	BB

Supervisor's Sign:

Name of the Research Supervisor: **Dr. Sudhir P Vegad**

ORIGINALITY REPORT CERTIFICATE

It is certified that PhD Thesis titled **Event based Video Summarization for Traffic Surveillance** by **Chauhan Ankita P.** has been examined by us.

We undertake the following:

a. Thesis has significant new work / knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

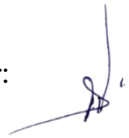
c. There is no fabrication of data or results which have been compiled / analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using **Turnitin** (copy of originality report attached) and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e. permitted similarity index <=10%).

Signature of the Research Scholar:  Date: 30th July 2026

Name of the Research Scholar: **Chauhan Ankita P.**

Signature of the Research Supervisor:  Date: 30th July 2026

Name of the Research Supervisor: **Dr. Sudhir P Vegad**

Place: Anand

Ankita Chauhan

199999913508 Event based Video Summarization for Traffic Surveillance

APC_7-7-2026

Research_SVEGAD_2025

Gujarat Technological University

Document Details

Submission ID

trn:oid:::1:3607285922

Submission Date

Jul 7, 2026, 2:32 AM GMT+5:30

Download Date

Jul 7, 2026, 2:38 AM GMT+5:30

File Name

199999913508_Chauhan_Ankita_Thesis-7-07-2026-Final.docx

File Size

31.7 MB

133 Pages

37,504 Words

224,154 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

294 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

4 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources

6% Publications

1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

PH.D. THESIS NON-EXCLUSIVE LICENSE TO GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a Research Scholar at Gujarat Technological University, and in the interests of the facilitation of research at the University and elsewhere, I, **Chauhan Ankita P.,** having (Enrollment No.) **199999913508** hereby grant a non- exclusive, royalty free and perpetual license to the University on the following terms:

a. The University is permitted to archive, reproduce and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as the "Work") anywhere in the world, for non- commercial purposes, in all forms of media;

b. The University is permitted to authorize, sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);

c. The University is authorized to submit the Work at any National / International Library, under the authority of their "Thesis Non-Exclusive License";

d. The Universal Copyright Notice (©) shall appear on all copies made under the authority of this license;

e. I undertake to submit my thesis, through my university, to any Library and Archives. Any abstract submitted with the thesis will be considered to form part of the thesis.

f. I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.

g. If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.

h. I understand that the responsibility for the matter as mentioned in the paragraph (g) rests with the authors / me solely. In no case shall GTU have any liability for any acts/ omissions / errors / copyright infringement from the publication of the said thesis or otherwise.

i. I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my university

in this non-exclusive license.

j. GTU logo shall not be used /printed in the book (in any manner whatsoever) being published or any promotional or marketing materials or any such similar documents.

k. The following statement shall be included appropriately and displayed prominently in the book or any material being published anywhere: "The content of the published work is part of the thesis submitted in partial fulfilment for the award of the degree of Ph.D. in Computer/ IT Engineering of the Gujarat Technological University".

l. I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my university in this nonexclusive license. I shall keep GTU indemnified from any and all claims from the Publisher(s) or any third parties at all times resulting or arising from the publishing or use or intended use of the book / such similar document or its contents.

m. I am aware of and agree to accept the conditions and regulations of Ph.D. including all policy matters related to authorship and plagiarism.

Date: 30th July 2026

Place: Anand

Signature of the Research Scholar:

Recommendation of the Supervisor: **Approved**

Signature of the Research Supervisor:

THESIS APPROVAL FORM

The viva-voce of the PhD Thesis submitted by **Chauhan Ankita P. (**Enrollment No. **199999913508)** entitled "**Event based Video Summarization for Traffic Surveillance"** was conducted on **Thursday, 30th July 2026** (Day and Date), at Gujarat Technological University.

(Please tick any one of the following options)

- [x] The performance of the candidate was satisfactory. We recommend that he/she be awarded the PhD degree.
- [ ] Any further modifications in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the Supervisor or request of Independent Research Scholar after which viva-voce can be re-conducted by the same panel again.

(Briefly specify the modifications suggested by the panel)

- [ ] The performance of the candidate was unsatisfactory. We recommend that he/she should not be awarded the PhD degree.

(The panel must give justifications for rejecting the research work)

Dr. Sudhir P Vegad
(Supervisor)

Dr. Sonali Patil
(External Examiner 1)

Dr. Gurucharansingh Sahani
(External Examiner 2)

ABSTRACT

Intelligent Transportation Systems (ITS), smart city infrastructures, and large-scale surveillance networks have created an ever-increasing amount of traffic video data. Today's traffic surveillance systems use a combination of CCTV cameras and roadside monitoring devices to monitor traffic continuously, which is inefficient, time-consuming, and unreliable for real-time traffic management when monitored manually. Human operators are only limited by cognitive fatigue, inconsistent decision-making, and lack of capacity to process extended surveillance streams. Rare events in traffic accidents are difficult to identify in continuous traffic videos due to a large amount of repetitive information in the video, high storage size, and high computational overhead.

There are several limitations in the current traffic accident detection and traffic video summarization methods in terms of scalability, number of computations, and real-time suitability. Most of the traditional machine learning and supervised deep learning models require large, labelled datasets that are costly and hard to gather in real-world surveillance scenarios. Many anomaly detection methods at the frame level are not capable of capturing temporal dependency and motion dynamics of complex traffic accidents. Most of the existing systems consider accident detection and video summarization as two different systems which leads to redundant computation, inefficient use of features, and high processing overheads, which is why an integrated unsupervised framework is proposed.

This research presents a comprehensive deep learning-based event-based video summarization solution in traffic surveillance with two unsupervised architectures based on reconstruction: Convolutional Variational Autoencoder for Accident Detection and Summarization (CVAE-ADS) and a Conv-BiLSTM autoencoder. The CVAE-ADS model is implemented as a probabilistic encoder-decoder and learns compact representations of normal traffic behaviour and detects anomalies by analysing the reconstruction error. Conv-BiLSTM Autoencoder model is a combination of Convolutional feature extraction followed by Bidirectional LSTM networks and aims to model spatial characteristics and temporal dependencies found in traffic video sequences. Both models are trained only with normal traffic scenes and can learn to detect anomalies without any labelled accident data.

The proposed framework is based on a sequence of processing stages, including video pre-processing, frame extraction, reconstruction-based anomaly detection, threshold-based classification, and summarization based on events. The surveillance videos are pre-processed by resizing and normalizing the videos, then input into the CVAE-ADS and Conv-BiLSTM models. Anomalous traffic events are detected by means of squared error reconstruction analysis, and frames that exceed the predefined threshold are sent to the summarization module. Images with anomalous frames are clustered together semantically using K-Means clustering, and the repetitive images are filtered out using the Structural Similarity Index Measure, resulting in concise and semantically meaningful summaries of the critical traffic incidents.

The proposed framework proves effective with experimental evaluation on the IIT Hyderabad (IITH) Road Accident Dataset and the UCF-Crime dataset. The Conv-BiLSTM Autoencoder outperformed on both the datasets for accident detection with an accuracy of 94.0% and 90.0% compared to that of conventional CAE and VAE-based methods. The CVAE-ADS framework successfully learned compact latent representations of the normal traffic behaviour with accuracies of 92.5% and 88.6%, respectively. The event-based summarization component was able to significantly reduce the redundant information within the video content (70–85%) while retaining the essential accident-related information.

The unified framework substantially enhances the computation efficiency, thus playing a major role in the smart city and ITS application of intelligent traffic surveillance. In summary, the proposed models can effectively and efficiently detect accidents and generate video summaries simultaneously and maintain a good compromise between the detection accuracy, compression performance, and reconstruction quality.

ACKNOWLEDGEMENT

Embarking on this research journey has been a path full of challenges, learning, and growth. I am truly humbled and grateful for the countless individuals whose guidance, encouragement, and support made this accomplishment possible. First and foremost, I bow my head in gratitude to the Almighty, whose blessings have been my constant companion throughout this journey. Without divine guidance and strength, none of this would have been achievable. I feel deeply blessed for the unwavering support that carried me through the long and sometimes overwhelming hours of research and writing.

It is with profound respect and sincere gratitude that I express my heartfelt thanks to my supervisor, **Dr. Sudhir P Vegad**, **Professor and Head, Department of Computer Engineering, G H Patel College of Engineering and Technology, Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar, Anand.** His invaluable guidance, insightful suggestions, and unwavering support have been instrumental in the successful completion of this research. His patience, encouragement, and dedication to academic excellence have inspired me throughout this journey. I am deeply grateful for the opportunity to work under his esteemed supervision, whose wisdom and mentorship have significantly enriched both this thesis and my academic growth.

I am equally thankful to the members of my Doctoral Progress Committee, **Dr. Narendra M Patel (Associate Professor, BVM Engineering College, Vallabh Vidyanagar, Anand)** and **Dr. Hiteshkumar M Nimbark (Provost, Gyanmanjari Innovative University, Bhavnagar)**, whose constructive feedback and steadfast encouragement have played an instrumental role in refining my work. Their sharp insights and expert suggestions have significantly elevated the quality of my research. I deeply appreciate their time, effort, and the academic rigor they brought to the table, which has shaped this thesis into what it is today.

I extend my sincere appreciation to my family, friends, and colleagues for their patience, understanding, encouragement, and belief in me throughout this journey. Their constant support has been a source of strength and motivation during the most demanding phases of this research.

I would like to express my deepest gratitude to my beloved parents, **Late Shri Pravinbhai Chauhan** and **Smt. Chandanben Chauhan**. My late father's values, vision, and blessings

have been a constant source of inspiration throughout my life. Together, my parents not only provided me with the gift of education but also instilled in me the values of hard work, perseverance, integrity, and humility. I am especially grateful to my mother, whose unconditional love, sacrifices, strength, and unwavering faith in me have shaped my character and guided me throughout my academic and personal journey.

I also extend my sincere thanks to my in-laws, **Shri Ranchhodbhai Parmar** and **Smt. Hansaben Parmar**, for their constant encouragement, understanding, and support. Their belief in my abilities and thoughtful guidance have provided me with the confidence and motivation to complete this research work.

To my beloved husband, **Mr. Rohit Parmar**, I offer my heartfelt gratitude. His sacrifices, unconditional love, patience, and unwavering support have been instrumental in enabling me to complete this research successfully. Through every challenge and achievement, he stood beside me with encouragement and confidence, making this journey both meaningful and rewarding. And to my sweet sons, **Darsh** and **Aaryav**, whose innocent smiles and boundless love have brought me immeasurable joy and support, this thesis is dedicated to you both, for being the light in the midst of all the hard work.

Finally, I extend my sincere thanks to everyone who has supported, encouraged, and believed in me, whether directly or indirectly. Your contributions, kindness, and goodwill have played an important role in this achievement. This acknowledgment is a small token of my immense gratitude to all who have been part of this journey. Without your collective support, this accomplishment would not have been possible.

Chauhan Ankita P.
Research Scholar,
Gujarat Technological University

Table of Contents

List of Abbreviations

Abbreviation	Full Form
5G	Fifth Generation Mobile Network
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under the Curve
A-VAE	Attention-based Variational Autoencoder
BiLSTM	Bidirectional Long Short-Term Memory
CADP	Context-Aware Dataset for Pedestrians and Vehicles
CAE	Convolutional Autoencoder
CCTV	Closed-Circuit Television
CNN	Convolutional Neural Network
Conv-BiLSTM	Convolutional Bidirectional Long Short-Term Memory
CVAE	Convolutional Variational Autoencoder
CVAE-ADS	Convolutional Variational Autoencoder for Accident Detection and Summarization
DDR4	Double Data Rate 4 (RAM type)
DETR	Detection Transformer
DL	Deep Learning
DOTA	Dataset for Object Detection in Aerial Images
EER	Equal Error Rate
ELBO	Evidence Lower Bound
EMA	Exponential Moving Average
FAR	False Acceptance Rate
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GBM	Gradient Boosting Machines
GPU	Graphics Processing Unit
IITH	Indian Institute of Technology Hyderabad
I/O	Input / Output
IoT	Internet of Things

ITS	Intelligent Transportation System
KL	Kullback–Leibler Divergence
k-NN	k-Nearest Neighbours
LSTM	Long Short-Term Memory
LTE	Long-Term Evolution (wireless communication)
MAD	Median Absolute Deviation
mAP	Mean Average Precision
MELM	Mixture of Extreme Learning Machines
ML	Machine Learning
MSE	Mean Squared Error
NVIDIA	NVIDIA Corporation (GPU Manufacturer)
PR	Precision–Recall
PSNR	Peak Signal-to-Noise Ratio
RAM	Random Access Memory
RCNN	Region-Based Convolutional Neural Network
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
RTX	Ray Tracing Extreme (NVIDIA GPU Series)
SHAP	SHapley Additive exPlanations
SSD	Solid-State Drive
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine
TAD	Traffic Accidents Detection Dataset
TN	True Negative
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbour Embedding
UCF	University of Central Florida
V2I	Vehicle-to-Infrastructure
VAE	Variational Autoencoder
VAID	Vehicle Accident and Incident Detection Dataset
VRAM	Video Random Access Memory
YOLO	You Only Look Once (Object Detection Algorithm)

Nomenclature

Symbol	Description
x	Input frame/image
$\hat{x}$	Reconstructed frame/image
z	Latent representation
μ	Mean of the latent distribution
σ	Standard deviation
σ^2	Variance
ϵ	Random noise sampled from a normal distribution
L_{rec}	Reconstruction loss
L_{KL}	Kullback-Leibler divergence loss
L_{total}	Total loss function
A_t	Anomaly score at time t
S_t	Smoothed anomaly score at time t
R_t	Reconstruction error at time t
T_{CVAE}	Statistical threshold for CVAE-ADS
T_{BiLSTM}	MAD-based threshold for Conv-BiLSTM
α	EMA smoothing factor
μ_{err}	Mean reconstruction error
σ_{err}	Standard deviation of reconstruction error
m	Median anomaly score
k	Threshold scaling parameter
x_t	Input frame at time t
z_t	Latent feature vector
h_t	Hidden state
F_t	Spatial feature representation
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

List of Figures

List of Tables

Chapter 1 Introduction

1. Introduction

The urbanization and the growing number of vehicles have intensified the difficulties in traffic monitoring, traffic management, and traffic safety. Over the last decade, large amounts of video surveillance data have been produced, and manual management is impractical due to the sheer volume. In addition to the difficulties mentioned, some minor but important events like traffic accidents and the extraction of meaningful information from continuous video streams are still considered research challenges.

This research aims to build an intelligent and unified system for traffic accident detection and video summarization based on deep learning. The proposed methodology is to combine both tasks into one system, which would provide efficient processing of surveillance data while avoiding duplication of content and computational load. The research will use unsupervised learning models and spatiotemporal feature extraction to improve the accuracy of the detection and generate a brief and informative summary to facilitate the decision-making process for intelligent transportation systems.

1.1 Traffic Surveillance and Intelligent Transportation Systems

Expansive urban growth has combined with exponential growth in the number of vehicles on the road to make traffic management systems in the world ever more complex. The urbanization process has resulted in extensive road networks, heavy traffic volumes, and growing demands for efficient modes of transportation. The effectiveness of the proposed framework is demonstrated through experimental evaluation on the IIT Hyderabad (IITH) Road Accident Dataset and the UCF-Crime dataset.

In order to solve these issues, a new paradigm, Intelligent Transportation Systems (ITS), which integrates communication technologies, sensors, and computation with transportation systems, is being adopted as a solution [1]. It integrates hardware and software elements like sensors, cameras, communication infrastructure, and data processing systems into one system to track, analyse, and control traffic conditions in real-time. The main goals of ITS are to optimize traffic flow, minimize congestion, minimize travel time, and increase the safety of the road by intelligent decision-making processes. ITS can be used to implement

real-time traffic control measures, predictive analysis, and automated incident detection, leading to enhanced transportation efficiency.

The traffic surveillance system is one of the most essential elements of ITS and is used extensively to acquire and monitor data. Traffic surveillance systems are a series of cameras and sensors that are placed in the urban road network, highway, and intersections and are constantly recording traffic activities [2]. Such systems provide enormous visual and sensor data, which are used to support traffic analysis, incident detection, and enforcement of traffic laws. Surveillance data can also be leveraged to provide vehicle flow monitoring, congestion detection, violation detection, and emergency response. Such real-time data has greatly improved authorities' ability to make decisions and take immediate action in the event of traffic incidents.

The combination of computer vision and deep learning techniques has reshaped traffic surveillance systems from passive data recorders to intelligent analysis systems in recent years [3]. The traditional surveillance systems mainly captured video data and then reviewed it after the event, where the review process was time consuming and error prone. Thanks to the progress of deep learning algorithms, especially CNN, however, it is now possible to automatically analyse video data and detect complex patterns and anomalies in real time. These systems are able to accurately sense many traffic events, including vehicle detection, vehicle classification, vehicle tracking and abnormal behaviour detection. This shift towards automating the analysis process has significantly reduced the amount of manual labour and improved the efficiency and accuracy of traffic monitoring systems.

Moreover, new technologies have also enhanced the capabilities of ITS, such as the communication network 5G and the concept of edge computing. High-speed data transmission with low latency has been achieved by the 5G technology, making it crucial for the real-time monitoring of traffic and decision making, which is required for the efficient transportation systems of the future [4]. Edge computing, however, enables data to be processed at the edge where the data is created, which helps to decrease the need for central processing and latency. This is especially crucial for traffic surveillance systems, as accurate and efficient recognition of the incident is essential. ITs can be made more efficient to handle vast quantities of data, more responsive and scalable with 5G and edge computing.

The integration of Internet of Things (IoT) devices and connected vehicle technology have significantly enhanced the capabilities of traffic management systems in tandem with communication technologies. The IoT frameworks can be used to integrate vehicles, sensors, and infrastructure, enabling data exchange and helping to enhance situational awareness [5]. Connected vehicles can share information about speed, location and traffic conditions with other vehicles and infrastructure in the immediate vicinity, thereby proactively managing traffic and preventing accidents. This environment is interconnected, and intelligent systems can be created that identifies potential issues and take preventative measures.

Another significant advancement in traffic surveillance is the emergence of spatiotemporal modelling methods which attempt to model traffic scenes, both in space and in time. Spatial features are features that are visual, such as the shape of the vehicles, and the structure of the roads, while temporal features are features that represent motion patterns and change over time. There are many models available to model temporal dependencies in videos, such as deep learning models, including recurrent neural networks (RNNs) or long short-term memory (LSTM) networks [6]. These models can be used to detect the similarities and differences among various traffic patterns, and consequently, they can be more accurate in detecting traffic accidents and identifying traffic anomalies, by merging both spatial and temporal information.

Despite these developments, there are still a number of obstacles to the effectiveness of traffic surveillance systems. The management of high volumes of video data from continuous surveillance systems is one of the difficulties. Such data need large computing power for storage, processing and analysis and efficient data management strategies. Moreover, detecting rare events such as traffic accidents is a challenging problem as diverse environmental conditions, such as weather, lighting conditions and occlusions exist. These factors have the potential to impact the performance of detection models and cause false alarms or missed detections.

In conclusion, with the development of other high-tech technologies such as deep learning, IoT, 5G and edge computing, the traffic surveillance systems and Intelligent Transportation Systems have undergone a tremendous shift. All these advances have contributed to the successful monitoring and control of traffic flows, enabling decision making, and enhancing road safety. Despite the progress made in data collection and processing, there are still significant challenges to address, such as the volume of data, computational complexity, and

the need for accurate detection of rare events. Further research and development are required to meet these challenges and create intelligent and efficient traffic surveillance solutions. Table 1.1 presents a summary of the development of traffic surveillance and accident detection methods from a manual monitoring system to advanced deep learning and Spatio-temporal systems. The table showcases the development of intelligent surveillance technologies, their main techniques, pros, cons, and the transition to automated traffic analysis systems.

Table 1.1: Development of Traffic Surveillance and Accident Detection Methods

Approach Type	Key Technique Used	Strengths	Limitations
Traditional Surveillance Methods	Manual monitoring and rule-based analysis	Simple implementation and low computational requirements	Time-consuming and error-prone
Machine Learning-Based Methods [6, 10]	SVM, Random Forest, feature-based classification	Adaptive learning and moderate accuracy	Requires handcrafted feature engineering
Deep Learning-Based Methods [5, 9, 13]	CNN, YOLO, Faster R-CNN, GAN-based learning	High detection accuracy and automated feature extraction	Limited temporal understanding
Spatio-Temporal Models [12,16]	CNN + LSTM / BiLSTM architectures	Captures temporal dependencies and motion dynamics	High computational complexity
IoT and Edge-Based Systems [1, 2, 7]	IoT sensors, connected vehicles, edge computing	Real-time processing and rapid incident response	Infrastructure dependency
Large-Scale Video Analytics Systems [8, 21]	Cross-modal learning and large-scale surveillance analytics	Efficient handling of massive video streams	Storage and computational overhead
Proposed Research Direction	CVAE-ADS and Conv-BiLSTM Autoencoder Framework	Integrated accident detection and summarization	Requires real-time optimization

The overall architecture of an Intelligent Transportation System (ITS), which features the co-operation of several components of traffic monitoring and management, is shown in

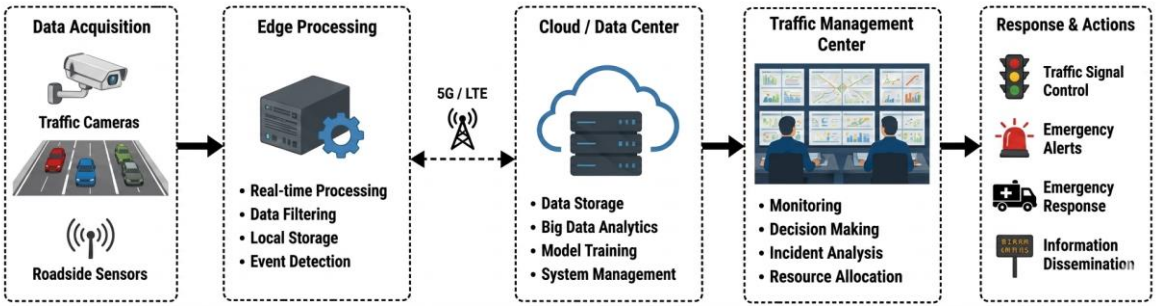


Figure 1.1: Architecture of an Intelligent Transportation System integrating data acquisition, edge processing, cloud computing, and centralized traffic management

Figure 1.1 Traffic data is gathered in the data acquisition layer, using traffic cameras and roadside sensors to collect traffic data from the road networks. They feature cameras, loop detectors, as well as environmental monitoring sensors that gather information about vehicles, traffic flow, and roads.

The data that is captured is then sent to the edge processing layer, where a preliminary analysis is conducted. By processing video streams in real-time, such as filtering data, extracting features, and detecting initial events, Edge computing helps lower latency and network bandwidth usage. This layer is very important for the applications that are time-sensitive, like accident detection. The data is then sent to the cloud or data centre using high-speed communication networks like 5G or LTE. The cloud layer delivers large-scale storage and advanced analytics, from model training and data mining to long-term analytics. It allows the system to process large amounts of data efficiently, as well as make intelligent decisions.

The information analysed is then used by the traffic management centre to monitor traffic conditions, analyse incidents, and make strategic decisions. Lastly, appropriate actions are implemented, such as traffic signal control, emergency warnings, and information dissemination, to ensure the safety and efficiency of the roads. The overall structure demonstrates the integration of different technologies, reflecting the comprehensive and intelligent nature of the traffic management solution.

1.2 Growth of Video Data and Need for Automation

With the rapid development of surveillance technology and the proliferation of camera-based surveillance systems in today's modern cities, the amount of video data created is growing at an unprecedented rate. Smart cities are home to thousands of surveillance cameras that are

used daily to solve traffic monitoring and safety problems on roads, intersections, highways, and public areas. They monitor round the clock and capture video footage of all sorts, producing a large amount of data every day. Thus, traffic surveillance systems are one of the largest producers of real-time data, which poses significant problems in the area of data storage, processing, and analysis [7].

Video data volume growth is fuelled by multiple factors, including camera technology advancements, high-definition video and ultra-high-definition video, and the proliferation of surveillance infrastructure. The modern cameras are able to give detailed visual data at high frame rates, meaning that they can offer higher quality surveillance, but also more data. Furthermore, the capacity to integrate data from many cameras across large geographic areas adds to that, increasing the amount of data that creates more problems for traditional data management.

Manual monitoring is one of the main problems involved with this data deluge. Traditionally, traffic surveillance systems depended on humans to monitor the video feeds and detect incidents. However, as more and more cameras and video feeds are deployed, manual monitoring is becoming impractical and inefficient. Humans can't keep track of events for long periods of time and can't maintain attention and avoid fatigue for long periods [8]. The more spread-out surveillance systems are, the more likely they are to overlook the event, which will result in a delay in response and an increased risk of exacerbating potentially critical situations such as traffic accidents.

An essential issue with manual surveillance is the variability in human decision-making. Different operators will see different traffic situations and will have different detection accuracy. Moreover, the number of events the operator has to observe and look for accidents is small compared to the number of normal traffic events, which means the operator must observe a lot of uneventful traffic events and search for a few critical events. This disproportionality is another factor that reduces the effectiveness of the manual monitoring mechanisms and underscores the importance of automated solutions.

To overcome these drawbacks, Automated Video Analysis systems have been developed using Machine Learning and Deep Learning. The ultimate aim of these systems is to process very large volume of video data and to be able to automatically generate useful information out of the video without using human intervention. The deep learning-based approach has

proved to be highly successful in visual information analysis, automatically learning and extracting hierarchical representations from raw data [9]. These can accurately identify objects, track vehicle movements, and identify unusual behaviour, improving the effectiveness of traffic monitoring systems.

Along with the deep learning models, cloud-based storage and processing power systems have been added to efficiently process the staggering amount of data generated by surveillance video. Cloud Computing makes it possible to store the data in a distributed manner and to process it in parallel, and then analyse it centrally, so that large-scale data can be processed without the need for a large amount of local computing [10]. This infrastructure also enables data sharing and remote access among multiple surveillance sites. With the need for real-time traffic monitoring and accident detection, though, the popularity of edge computing architectures has been rising. By processing data near the data generation, edge computing helps to reduce latency, bandwidth usage, and response time. Cloud/Edge computing solutions provide scalable and efficient solutions that can be used for intelligent traffic surveillance and transportation applications.

With these developments, however, there are some technical challenges to transition from manual to automated surveillance. The processing of high-resolution video streams must be done promptly on high performance computers, such as GPUs, and with optimized deep learning algorithms. With the growing number of cameras and increasing volume of video data, ensuring scalability and real-time performance becomes challenging. Moreover, the continuous surveillance systems may also produce a lot of redundant information, with redundant frames adding to the overhead of storage and processing without significantly adding value. Thus, the use of effective data reduction and event extraction techniques are important in improving the system performance. In this context, video summarization techniques are relevant since they are able to extract keyframes and important events, discarding redundant information and retaining significant traffic information that may be analysed and applied in decision making.

Furthermore, environmental issues such as illumination changes, bad weather and occlusion cause another challenge in video analysis. All these variations and different conditions must be taken into account, and the automated system must be robust enough to deal with all of them and perform the same. A very strong performance level which requires advanced level of model design and extensive training. This need for automation in traffic surveillance is

thus a consequence of the disadvantages of manual monitoring, the growing amount of video data, and the need for real-time and accurate analysis. Automated systems can help to increase efficiency and provide scalable solutions to meet the demands of today's complex transportation systems.

As mentioned in the introduction, the manual method of traffic monitoring has its own drawbacks, and when compared with the automated method, can be seen in Table 1.2 comparative performances of these two types of monitoring systems. A manual surveillance system is user-limited, has poor scalability, and is prone to inconsistent decision-making, fatigue, and attention spans. Automated systems use deep learning and cloud-edge computing to constantly monitor and process video data in real time, improve accuracy, and facilitate the effective management of video data, especially when dealing with massive amounts. The comparison emphasizes the rising need for smart automated surveillance solutions in today's traffic management landscape and their role in enhancing road safety.

Table 1.2: Manual Traffic Surveillance System vs. Automated Traffic Surveillance System

Feature	Manual Surveillance	Automated Surveillance
Monitoring Capability	Limited by human attention	Continuous and scalable [9, 10]
Accuracy	Inconsistent and error-prone	High accuracy with trained models [9]
Response Time	Delayed response	Real-time or near real-time processing [7, 10]
Scalability	Poor scalability	Highly scalable with cloud-edge infrastructure [10]
Data Processing	Manual and slow	Automated and fast
Handling Large Data	Inefficient for large-scale video streams	Efficient large-scale video analytics [7, 10]
Fatigue Factor	High human fatigue and inconsistency	No fatigue-related performance degradation

Cost Efficiency	High operational and manpower cost	Lower long-term operational cost

Finally, video data has been increasing exponentially in traffic surveillance systems, making it difficult to achieve this manually. In this context, automated video analysis systems, supported by deep learning, cloud, and edge computing, could provide a potential solution to more efficiently manage large quantity of data. But computational complexity, redundant data, and real-time processing are important issues that need to be explored. To tackle these challenges, it is crucial to design and build intelligent and scalable traffic surveillance systems that can significantly contribute to improving road safety and traffic management in today's cities.

1.3 Accident Detection in Traffic Surveillance

Road traffic collisions are among the most important issues in contemporary transportation systems that annually cause millions of deaths, injuries, and economic losses worldwide. As the number of vehicles on the road grows, and traffic patterns become increasingly complex, along with human error, the occurrence of accidents increases, and road safety is an increasing concern to governments and transportation agencies. In this aspect, the capacity of good accident detection systems to minimize the impact of an accident through timely interventions, improved accident response activities and provision of traffic management strategies, is important [11]. An immediate notification of the accident can help to minimise response time to recovery operations, reduce the risk of a secondary collision and enhance normal traffic flow restoration.

Traffic monitoring systems are crucial to accident detection: cameras are deployed to stream video of traffic on the road and are used to constantly monitor traffic. These systems can deliver live visualization of the data, which can then be analysed to detect any abnormal incidents or dangers. The manual monitoring is limited and, together with the huge amount of video data in current traffic systems, the automatic analysis of the video and recognition of scenes of accidents becomes increasingly necessary. The objective of the automated accident detection systems is to be able to process these video streams efficiently and find unusual patterns that might signal an accident has occurred.

Accident detection is primarily the process of detecting abnormal traffic conditions from normal. These deviations can range from sudden crashes and abrupt braking to irregular courses and unexpected stops. The rare events are in general rare in terms of the traffic volume (when compared to the normal traffic conditions), and they are often detected under varying and unpredictable traffic conditions, which makes the detection task difficult [12]. The variety of traffic conditions, which includes variation in the roadway structure, type of vehicles, light condition, and weather condition, further complicates the problem of accurately identifying the accident events.

The conventional approaches for accident detection relied on rule based and handcrafted feature extraction techniques. These systems were based on handcrafted rules and features such as motion vectors, object shape, speed threshold etc. that were used to detect anomalies. The approaches could perform simple operations but could not be adapted and could not deal with complex and changing traffic situations in the real world. They were not performing well in the real world and were not easily adapted to the dynamic environments in which they would be used and were often found to have low detection accuracy and high false alarm rates.

With recent advancements, the development of DL based techniques has greatly improved the effectiveness of accident detection frameworks, enabling efficient feature learning and robust pattern recognition. CNNs have been widely used to extract spatial information from the frames in the video, such as shape, position, and context, of the vehicles under observation [13]. These models can be applied to learn a hierarchical representation of visual data, which can help them in identifying complex patterns in the context of accident events. Furthermore, deep learning methods do not require much feature engineering, which makes them more scalable and flexible to various traffic scenarios.

Unsupervised learning models have attracted much research interest in the field of accident detection among DL based approaches. The models do not need labelled data sets, which are challenging and costly to acquire in the real world. In this context, the autoencoder-based models are more widely used because they learn to reconstruct normal traffic patterns during training. If an abnormal event happens, for example, an accident, because the model cannot reconstruct the abnormal input, it will have a higher reconstruction error [14]. This reconstruction error is used as an indication to identify anomalies in surveillance videos.

Although the models based on autoencoders have some pros, they also have their drawbacks. One of the main problems is the lack of the ability to effectively represent temporal dependencies between the two consecutive video frames. Motion patterns play an essential role in accurate traffic accident detection, as traffic accidents are dynamic events. Temporal context may be disregarded by models that are only spatial, resulting in events being misclassified or not detected [15].

To overcome this limitation, a hybrid CNN and RNN (e.g., LSTM) have been proposed. In traffic analysis, these models are used in combination with spatiotemporal feature extraction and sequence modelling. The hybrid architectures can only capture both spatial and temporal information, which can lead to better identification of complex accident scenarios and higher detection accuracy [16]. For example, CNN layers extract the visual features of every single frame, whereas LSTM layers analyse the temporal evolution of the features over time, and the model is able to recognize the patterns related to accidents.

These models have come a long way with regard to accident detection, but there are still several challenges. Some environmental problems, such as lighting, weather, shadows, and occlusion, can have a negative impact on the model's accuracy. For example, during nighttime, when the background lighting may be weak or when it is raining, when the visual details are not easily discernible from the model, the model is unable to detect anomalies. Similarly, there may be other vehicles or objects that block a particular piece of information, thus leading to misjudgements. A big problem is also the case of false alarms, in which typical situations are judged as being abnormal. High false alarms can lead to a loss of confidence in the system's integrity, and they can overwhelm traffic officials.

Real-time deployment can also be difficult, particularly in large-scale surveillance systems with multiple video streams, as the complexity of advanced models can be challenging. Furthermore, the traffic conditions will change significantly from one location to another, making it difficult to develop a model that is applicable to all these conditions. It is important to have strong and flexible models because these models may not work well when deployed in a different environment for which they were not created.

In summary, the detection of moving objects in traffic surveillance systems is a complex but vital task for enhancing road safety and traffic control. Despite the success of DL based methods in boosting the performance of detection systems, temporal modelling,

environmental variations, and computational efficiency remain issues. To address these challenges, there is a need for the development of more powerful, and intelligent models with the capability to accurately perceive the accident events in real-world scenarios.

The overall structure of an automated traffic accident detection system using deep learning techniques is given in Figure 1.2. The first step is the input video stream, which is the continuous video footage from the traffic surveillance cameras. The video is broken down into frames here, so that each frame can be analysed individually.

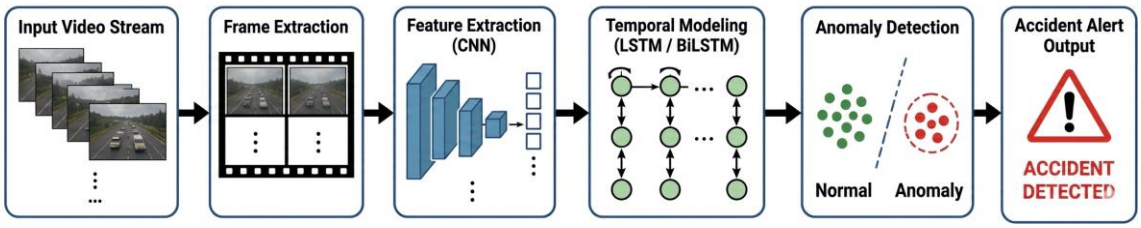


Figure 1.2: General framework for traffic accident detection using spatial and temporal feature learning from surveillance video streams

A feature extraction module, commonly based on CNN, is then used to extract the features from the extracted frames, which are then passed to the next module. This module is concerned. With identifying spatial elements such as the shape of vehicles, spatial patterns of motion, and spatial contexts. These features give valuable data to understand the visual properties of traffic scenes. After spatial feature extraction, temporal modelling is done using RNN approaches like LSTM and Bidirectional LSTM (BiLSTM). In this stage, the temporal dependencies and motion dynamics between the consecutive frames are captured, thus enabling the system to analyse the changes in traffic behaviour over time. In particular, the temporal modelling is crucial for the detection of dynamic events like collisions and sudden motions.

In the anomaly detection stage, the features derived from the processed traffic are analysed to identify normal and abnormal traffic patterns. Once an anomaly is detected, which is indicative of a potential accident, the system sends an alert output showing that an accident has occurred. The automated pipeline can detect traffic incidents, minimizing response time and enhancing the safety of the road.

1.4 Event-based Video Summarization

With the increasing amount of video data captured by surveillance systems, video summarization has become a crucial technique for handling and understanding these massive amounts of data. Traffic surveillance systems record long video streams, which have a lot of redundant and irrelevant information. This is because it takes a long time and is costly to process and analyse these enormous amounts of information in its raw state. Video summarization is the process of extracting the important and relevant parts of the video data with reducing the complexity and retaining the critical information. The specific purposes of event-based video summarization in traffic surveillance are to detect and summarize the major events like accidents, traffic violations and unusual traffic interventions, and to remove redundant frames that contain no useful information [17].

The main goal of video summarization is to aid in rapid understanding and storage of surveillance information. Summarised outputs (including key events) allow operators to save time reviewing lengthy video recordings and make informed decisions and take prompt actions. This capability is particularly important in traffic monitoring applications, where timely identification of incidents is crucial for ensuring road safety and effective traffic management [18]. Summarization methods can also be used to reduce irrelevant information and quickly retrieve data.

The traditional video summarization methods are primarily low-level visual based methods, which rely on motion intensity, colour distribution and saliency detection. Though these approaches are feasible in computation, they cannot explain the meaning of complex events, especially in the dynamic traffic situation. This can cause some events to be missed or for irrelevant frames to be added to the summary. To overcome the limitations, deep learning approaches can be used to learn high level semantic features and temporal relationship from the video data [19]. The techniques used in them are complex neural network models that are able to capture context to generate more informative and relevant summaries.

The majority of the event-based summarization methods employ feature extraction and clustering based keyframes selection techniques for video sequences. These techniques ensure that the summaries created are brief and informative by clustering the similar frames according to feature similarity and choosing representative frames [20]. Not only do these techniques remove redundant information, but they also preserve the important information

of key events that are pertinent to a variety of uses such as traffic incident analysis and forensics.

One of the major disadvantages of current video summarization techniques, however, is that they are still regarded as stand-alone techniques not integrated with event detection system. This separation causes inefficiencies, with detection and summarization tasks potentially duplicating computational work, and not sharing useful information. So, the context of the events detected may not be fully summarized. The limitation suggests that a joint framework is required to combine accident detection with video summarization in one framework to make efficient processing and better representation of the critical traffic using the summarized video.

1.5 Motivation for Integrated Detection and Summarization

Existing approaches split the two functions of accident detection and video summarization, leading to the waste of computation and efficiency loss. Detection systems are used for detecting anomalies and summarization systems process video data independently to produce summaries [21]. When information is not integrated, there is a computational burden, and system performance does decrease. This can be achieved more efficiently by using a combined detection and summarisation system, which can share intermediate representations, avoiding redundant processing, as much as possible [22]. Such accidents can be detected by this type of system, and a meaningful summary can be given in one pipeline. Another dimension of scalability is gained by unsupervised learning techniques, which do not require labelled data sets, which can be scarce in practical applications [23]. What's more, spatial and temporal modelling can be combined to enhance the system's capacity to comprehend complex traffic dynamics [24]. The motivation of this research is to find an efficient solution for traffic surveillance, and such issues need a unified framework.

1.6 Problem Statement and Research Questions

The traffic surveillance system has developed rapidly, and the amount of video data generated has reached a huge scale, which makes it hard to monitor video manually. The real difficulty of using continuous video streams to detect rare accident events is that they are unpredictable, unpredictable traffic flows, and variations in the environment, including occlusion, lighting conditions, and weather disturbances. The existing methods typically fail

to correctly detect such anomalies, particularly when they do not take temporal dependency into consideration and rely on frame-level analysis [25, 27].

Moreover, in most existing systems, these two tasks of recognition of accidents and video summarization are performed separately. This separation will cause redundancy in processing, high computing costs, and inefficiency in feature extraction. Consequently, these systems are incapable of providing the full system solution needed to detect accident events and produce meaningful summaries to quickly interpret and decide. To tackle these challenges, this research aims to create a common framework that can effectively detect and summarise text while being scalable and efficient [29, 30].

Research Questions:

- How to apply unsupervised deep learning techniques to effectively detect traffic accidents in a scalable and adaptable manner without the need for large-scale labelled datasets in the real world?
- How to model both spatial (appearance-based) and temporal (motion-based) features together to better detect complex and dynamic traffic anomalies like accidents?
- What is the best way to automatically recognize and summarize the most useful event information from long surveillance videos without losing important frames from accidents and discarding redundant and irrelevant information?
- How do you minimize redundancy in continuous video without compromising the meaning and content of key events in the processed video?
- How to combine the above two tasks in a single unified system without compromising the efficiency, computational cost, and overall system performance?

1.7 Objectives, Scope and Outcomes

The overall objective of this research is to develop an efficient and uniform deep learning-based architecture for traffic accident detection and summarization of video data through a traffic accident event. The advantage of the proposed framework is that it tries to combine the detection and summarization processes in one pipeline to overcome some of the limitations of the existing systems and make the whole system more efficient, less redundant, and more applicable to real-time use in intelligent transportation systems.

Objectives:

The objectives of this research are as follows in detail:

- To build an unsupervised framework for detecting anomalies: Develop a model that can learn normal traffic patterns from surveillance videos and detect accidents based on the deviation from the learnt traffic pattern, without requiring labelled data to make the model with potential for practical application.
- To effectively capture spatial features: Apply CNNs to learn useful spatial information like vehicle shapes, patterns, and scene context from individual frames.
- To add temporal feature modelling: Explore and build models to learn temporal dependencies and sequential information from video data with advanced architectures, such as BiLSTM, to understand traffic dynamic behaviour.
- To develop a video summarization mechanism that is based on events: To obtain the key representative frames to retain essential information from accident events detected, propose a summarization method based on a latent space clustering technique.
- To minimize the duplication of video information: Use similarity-based-filtering techniques (e.g., SSIM) to drop redundant frames and make efficient summaries without losing important event information.
- To integrate detection and summarization in one structure: Improve the computational efficiency, processing time, and system performance by putting both processes in a pipeline.
- To assess the system's performance, using common performance measures: Compare and assess the proposed framework using evaluation metrics like Accuracy, Precision, Recall, and F1-score for Accident detection, Diversity and Reduction Rate for summarization.

Scope:

This research aims to concentrate on the intelligent traffic surveillance system using video and deep learning technology. This research deals with analysing the CCTV footage for the identification of accident events and the generation of meaningful summaries. The proposed framework is examined using benchmark datasets like UCF-Crime and IITH under a real-world traffic environment. The study underscores the importance of unsupervised learning techniques, which aim to minimize the need for labelled data and enhance scalability. It also

takes into account spatial and temporal modelling approaches for the representation of complex traffic dynamics. Furthermore, the study will investigate the computational efficiency and assess its feasibility for a practical application in smart city environments.

Expected Outcomes:

The research results of such sort are both theoretically and practically important. The suggested framework is expected to be highly accurate in detecting accidents, as it will be able to capture the spatial and temporal patterns in traffic videos. The system will use unsupervised learning methods, which should allow it to transfer its knowledge to varying traffic situations with fewer labelled data sets. Moreover, the event-based video summarization functionality is anticipated to provide the user with concise and informative video summaries by choosing the most representative keyframes while removing redundant frames. This will result in a high reduction rate in video data without compromising the quality and relevance of important events. The summaries will help traffic and emergency response authorities to analyse and make decisions more quickly.

Another major benefit will be the merging of detection and summarisation to one, thereby decreasing the computational burden and increasing the efficiency of processing. The integration could make the system more applicable to intelligent transportation systems. Overall, the study is expected to contribute significantly to the field of traffic surveillance, offering insights and potential solutions for the Advancement of efficient and intelligent systems to enhance safety on the road and traffic management.

1.8 Thesis Organization

This thesis is systematic and structured; the development and organization are intended to provide a full understanding of event-based video summarization and anomaly detection for an intelligent traffic surveillance system. Then, the basic concepts of traffic surveillance, ITS, accident detection, and event-based video summarization are introduced in Chapter 1. It explains the motivation behind combining detection and summarization in one single system, outlines limitations of existing surveillance systems, and outlines the problem statement, research objective, scope, and expected outcomes of the proposed research.

The literature review of traffic surveillance, accident detection methods, video summarization methods, spatiotemporal modelling methods, and Autoencoder-based

anomaly detection methods is discussed in Chapter 2. This chapter discusses and critically analyses the traditional surveillance methods, the ML-based technique, the DL based technique, event-driven surveillance methods, and their limitations, comparative study, and research gaps. The theoretical and practical support for the proposed integrated surveillance framework is provided in this chapter.

The theoretical background and basic concepts for the proposed research work are discussed in Chapter 3. It introduces you to the features of surveillance videos, the basics of deep learning, convolutional neural networks, autoencoder architectures, temporal modelling methods, clustering methods, similarity measures, and metrics for measuring the performance of autoencoders. The chapter offers the theoretical underpinning of the suggested CVAE-ADS and Conv-BiLSTM frameworks.

The integrated traffic anomaly detection and incident-based video summarization proposed models and overall system architecture are detailed in Chapter 4. The chapter introduces the design and implementation of the CVAE-ADS framework and Conv-BiLSTM autoencoder framework, which comprise a variety of preprocessing strategies, encoder-decoder architectures, latent feature representation, anomaly detection mechanisms, temporal modelling, thresholding techniques, and intelligent video summarization methods. It also provides a detailed description of the complete algorithmic flow of the proposed system.

The experimental setup, descriptions of the datasets, the preprocessing techniques, baseline models, and the metrics employed for the assessment of the systems are provided in Chapter 5. The chapter contains full quantitative and qualitative evaluations of the accuracy of the accident detection and summarization of the videos from the IITH and UCF-Crime datasets. Comparative studies with existing state-of-the-art techniques are also given in order to show the strength and effectiveness of the proposed frameworks.

Lastly, Chapter 6 presents the overall conclusions, research contributions, and main findings of the proposed work. The chapter also examines the drawbacks of the existing frameworks and proposes future research avenues on intelligent traffic surveillance, anomaly detection, and event-based video summarization. This structured organization facilitates understanding of the proposed integrated traffic surveillance system.

Chapter 2 Literature Review

2. Literature Review

The literature review provides an extensive analysis of the past research on spatiotemporal modelling, video summarization, traffic surveillance, accident identification, and anomaly detection techniques. It examines the evolution of traditional and intelligent systems for traffic monitoring and the significance of the role played by ML and DL-based techniques in replacing the traditional image processing techniques. The survey also includes recent advances in using autoencoders for anomaly detection and generating video summarises for effective processing of traffic surveillance data. To articulate the need for, and rationale behind, the proposed framework, a comparative evaluation of existing frameworks is also included to identify the strengths and weaknesses of current frameworks and research needs.

2.1 Overview of Traffic Surveillance Systems

The traffic surveillance systems are a vital component of today's transportation system and play a critical role in real world traffic analysis, monitoring and management. The number of vehicles has been increasing and so has the complexity of urban transportation, thus the traditional methods of traffic monitoring are no longer sufficient to the modern traffic managements. In this way, the traffic surveillance system has been transformed into a smart and automated platform, which is equipped with advanced sensing, communication and computational capabilities, so as to effectively manage the traffic flow and enhance the road safety [31].

The goal of traffic surveillance systems is to collect, process, and analyse traffic surveillance data from various sources like dashcams, surveillance cameras, connected vehicles, and sensors. Today, traffic surveillance systems are multi-layered systems that consist of data acquisition, data processing, and decision-making parts. Data acquisition is also installed on road networks, including cameras and sensors that continuously gather traffic data. This data is then fed into more sophisticated algorithms to analyse and derive significant features and patterns. Lastly, the processed information is being used to control traffic, manage incidents, and enforce policies [32].

The advent of big data and computer technology has also greatly helped the creation of more complex surveillance systems. Real-time monitoring of high-density and mixed traffic

conditions is possible with systems like TRAMON, which are able to merge real-time data from various sources [33]. Likewise, technologies such as Auto track allow effective vehicle monitoring and retrieval from surveillance videos, thereby improving the performance of traffic monitoring systems [34]. The developments indicate a move from passive to pro-active and intelligent traffic surveillance systems with the capability to make decisions, such as the use of event-driven video analytics [35]. But some traffic surveillance system challenges remain, including processing huge video data, real-time processing, and accuracy in varying environmental conditions. These tasks can only be achieved with continual innovation of hardware and software technologies and development of robust and efficient models.

2.1.1 Evolution of Traffic Monitoring Techniques

The need for effective traffic management and technological capabilities has influenced the development of traffic monitoring technology. The first approach to traffic monitoring was to have human monitors observing the traffic on the road using a camera or by sight. While this gave the basic monitoring, it had the drawback that it was subject to human factors such as fatigue, limited attention span, and was not scalable to traffic load.

The next step was the use of sensor-based systems, bringing automation to the collection of data. The traffic parameters measured by different technologies, such as the number of vehicles moving in the street, their speed, and vehicle occupancy, included inductive loop detectors, radar devices, and infrared devices. The use of these systems improved traffic data collection with much greater accuracy and reliability. However, they were unable to offer granular information on traffic scenarios or to analyse complex scenarios.

As communication technology has developed and the number of mobile devices is increasing, a new method of traffic monitoring has been introduced: vehicular crowdsensing. This approach involves the use of sensors and communication technologies to transform vehicles into mobile data collection and transmission sources, providing real-time traffic information. Previous research has shown that vehicular crowdsensing can be used to identify traffic accidents and enhance coverage in a large-scale transportation network [31]. This will improve the amount of data available, as well as provide a fuller picture of traffic conditions.

Advanced deep learning models like CNN, YOLO based models, RNN, and Vision Transformer have been widely used in the field of traffic analysis. Such models can be used to learn hierarchical representations of data and to find subtle patterns and anomalies that are hard to find otherwise. Along with that, some attention-based schemes and hybrid architectures are also implemented to achieve better efficiency of the ITS.

Another big trend in recent times is the implementation of edge computing. It can reduce latency and enhance efficiency by processing data on the edge. It is especially crucial in applications like accident detection, where a timely response is necessary. The success of systems like VegaEdge in providing the edge architecture capabilities to enable real-time IoT-based applications for highway safety has been shown [36]. Further, the real-time video analytics used on the edge/end devices have enhanced the monitoring of urban safety [37].

Some special applications include smoker detection in smart city [91], detection of the number plate of the vehicle in the video [92], near-infrared road-marking detection [93], and Motorcycle Helmet Detection in Traffic Surveillance [94]. In conclusion, the development of traffic monitoring methods has seen a gradual transition from traditional and reactive monitoring methods to intelligent and proactive systems. This shift is the result of the current demand for scalability, accuracy, and performance in today's traffic monitoring systems.

2.1.2 Conventional Surveillance Methods

Conventional surveillance methods are the initial solutions for traffic monitoring and incident detection, which are mostly rule-based, statistical, and basic tracking technologies. In a relatively controlled environment, these techniques were meant to take place under rules and thresholds already set forth to detect abnormal traffic situations. A technique widely adopted for the conventional surveillance systems is statistical modelling of the urban traffic flow [38]. There is also an attempt to use traffic analytics for low-frame-rate videos to deal with the bandwidth problem [39]. One way is using traffic flow models that are used to analyse traffic parameters like speed, density, and flow rate, and sense anomalies. Dynamic traffic signal control has been suggested for controlling urban congestion by using smart traffic surveillance and accident detection systems [40]. Although these models offer a structured approach to analysis, they rely on a set of assumptions that may not be applicable to other traffic situations and are not necessarily representative.

Another conventional approach is to detect the presence of people in video surveillance using artificial intelligence to ensure pedestrian safety [41]. The moving vehicle detection using the machine learning approach has also been applied to enhance road safety [42]. While effective in some scenarios, they are not always able to process complex traffic behaviour and identify subtle abnormalities as well as modern deep learning alternatives. Apart from that, manual feature extraction approaches were used with early surveillance systems, in which the features were designed according to the knowledge of the domain experts. The features were then utilized in classification or detection models. This, however, is a time-consuming process and does not fully represent the diversity and complexity of traffic situations on the road.

The problem with conventional surveillance systems is that they are not as flexible and can't adapt to the shifting dynamics. Weather, lighting, road structure, driver behaviours, and all other factors can have an impact on traffic conditions. The difficult variability is a problem for rule-based systems and results in lower accuracy and false alarms. Moreover, these approaches are not scalable and are unable to handle large amounts of video data produced by today's surveillance systems. In spite of these drawbacks, traditional surveillance techniques pave the way for more sophisticated techniques. They gave us a lot of insights into traffic behaviour and the need for more robust and adaptive systems.

2.2 Accident Detection Approaches

The detection of accidents is a key aspect of intelligent traffic surveillance systems that not only directly affects road safety but also contributes significantly to emergency response and traffic management efficiency. As the number of vehicles on the road continues to rise and the complexity of the road network grows, it is more important than ever to have accurate and real-time accident detection systems. In the literature, there are many methods proposed for the detection of traffic accidents, including traditional rule-based systems, as well as deep learning-based systems.

The main goal of accident detection systems is to detect unusual traffic incidents, including collisions, rapid stops, and unusual movements of vehicles, and then report them from the surveillance data. These events are often unusual and unpredictable and are hard to detect. In the past few years, research efforts have been directed towards improving the detection accuracy, minimizing the number of false alarms, and facilitating real-time performance by

developing better algorithms and models. Generally, there are three categories of accident detection methods: Conventional methods, Machine learning, and Deep Learning techniques. While they all have their pros and cons, the type of method used will vary according to the data available, the amount of computational resources required, and the type of application.

2.2.1 Traditional Methods

The traditional approach to accident detection is mostly rule-based logic, statistical analysis, and mathematical modelling of traffic parameters. These methods are based on rules, thresholds, and handcrafted features, which are pre-defined and used to detect abnormal traffic behaviour that could signal the presence of a road accident. Traditional Traffic surveillance systems were more concerned with analysing the vehicle speed, traffic flow, vehicle density, occupancy, and sudden changes in movements to identify unusual events. The system created an accident alert if the traffic condition was found to be very different from the normal condition.

Fuzzy logic-based systems are one of the popular methods used in the traditional accident detection system. These techniques are useful, especially for coping with uncertainty and imprecise traffic conditions by using linguistic rules and membership functions. Bhanja et al. [43] introduced a fuzzy logic accident detection system for VANETs in which they used the differences in vehicles' speed and communication signals to detect accidents. In the work of SP et al. [44], the authors have proposed a smart junction framework using intelligent traffic control and anomaly detection techniques to enhance road safety and traffic efficiency. Likewise, fuzzy Bayesian techniques have also been used to assess risks of accidents in transportation systems. In traffic safety analysis, Benallou et al. [45] validated the use of probabilistic reasoning in analysing the risk of accidents for road transportation providers and truck drivers.

Infrastructure-assisted and traffic-control-based systems are other traditional approaches. What would be the impact on accident detection if there were improved communications between vehicles and roadside infrastructure? Tian et al. [46] created an accident detection framework for cars automatically that leverages cooperative intelligent transportation frameworks. Besides that, an advanced zone-based traffic management system integrating with anomaly detectors has been proposed to improve urban traffic monitoring. In addition,

Zhu et al. [47] presented a traffic accident detection algorithm utilizing comprehensive traffic flow feature extraction, which was able to combine several traffic flow characteristics to enhance the detection performance.

Though the traditional methods are simple enough and computationally efficient, there are certain limitations associated with them. They are heavily dependent on pre-defined rules and hand-crafted features and are unable to cope well with the complex and dynamic traffic situations. Furthermore, these methods tend not to be accurate in detecting subtle interactions between vehicles, which results in lower detection accuracy in real-world situations.

2.2.2 Machine Learning-Based Approaches

The machine learning based approaches are a major improvement over conventional approaches as they allow data-driven modelling of traffic behaviour. They are based on algorithms that are trained with the historical data and then used to identify anomalies in the new information. This model can dynamically adapt to traffic situations and get better with time unlike rule-based models. Several ML techniques have been used in the field of accident detection, such as Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Gradient Boosting Machines (GBM), and Random Forests [26, 42]. Usually, such models consist of two steps, namely feature extraction and classification. The characteristics of the vehicles, such as their speed, trajectory, motion patterns, and the characteristics of the traffic, are extracted from the surveillance data and serve as features to train classification models. These studies have demonstrated that machine learning approaches can be successful in processing large amounts of traffic data and detecting complex patterns that might not be easily spotted through traditional means [46].

But some drawbacks for machine learning techniques exist as well. One of the big hurdles is the need for labelled datasets, which are often hard to get hold of and costly. Annotating traffic video data for detecting accidents is an expert skill and time-consuming. Moreover, the quality of the features extracted is a crucial factor in machine learning models. If care is not taken, inadequate feature selection may result in low accuracy and performance. One of the other drawbacks is that many machine learning models are not able to learn temporal dependency in video data. Traffic collisions are dynamic events that change over time, and the analysis of individual frames, without taking temporal relationships into account, may

lead to false predictions. This has resulted in the development of more sophisticated models incorporating time information [42, 47].

2.2.3 Deep Learning-Based Methods

The automatic feature extraction and modelling of complex spatial and temporal traffic patterns using deep learning models can be achieved in surveillance video and can greatly improve the capability of traffic accident detection. DL models can learn hierarchical representations directly from the raw images and video sequences, unlike the traditional methods, which rely on hand-crafted features, to deliver superior accuracy and robustness. Recently, deep learning and transfer learning methods have been useful in accident analysis and detection. Aboulola [48] proposed MobileNet using a SHAP-based explainable AI approach for the traffic accident severity prediction. Kaur and Bhattacharya [49] proposed a convolutional deep network for automatic traffic detection and classification. In addition, Pashaei et al. [50] presented a hybrid approach consisting of CNNs and a Mixture of Extreme Learning Machines (MELM), which enabled better feature extraction and classification accuracy of accident images.

CNNs are widely used for the applications of extracting spatial features such as vehicle appearance, motion features, collision patterns, and scene context. CNN combined with Variational Autoencoders (VAEs) has been proven to be a good architecture for traffic anomaly detection and vehicular trajectory classification tasks [51]. Moreover, the development of large-scale benchmark datasets, such as the Traffic Accidents Detection (TAD) dataset, which offers video surveillance with annotated data for training and testing models, has further boosted research in this field [52].

Recurrent and spatiotemporal learning frameworks are often combined in deep learning approaches to analyse temporal dependencies in traffic videos. Accident detection based on video is more effective in detecting accidents using sequential frame information and traffic motion patterns [53]. Furthermore, vision-based accident anticipation systems are supposed to enhance road safety by anticipating possible accident situations in advance [54]. More advanced anomaly detection techniques using optimized neural networks and spatiotemporal feature extraction techniques have been applied to improve the detection of abnormal events in surveillance scenarios [55]. Moreover, deep ensemble models have been proposed to deal with various traffic conditions and improve the models' robustness and generalization [56].

In recent years, the emphasis has been on accident detection techniques through advanced object detection and tracking systems. CCTV system-based intelligent accident detection system combined with a nearest medical facility recommendation system is used for applications in emergency response management [57]. The use of scenario-based detection frameworks, such as YoFlow, enhances the understanding of the roadside scene and the localization of an accident [58]. The researchers have successfully created advanced learning approaches with high detection accuracy and speed using Faster R-CNN, 3D-CNNs, YOLOV5, YOLOV8, and Generative Adversarial Networks (GANs) for intelligent traffic monitoring systems [59, 60, 68, 69, 71]. Trajectory tracking and influence map-based methods have also been used to analyse vehicle interactions and collision risks [61], and the computer vision-based approaches have been used to classify traffic incidents automatically from surveillance streams [62].

Moreover, synthetic accident videos have been employed to address the issue of a lack of real-world training videos [63]. With the advent of edge computing and deep neural network-based technologies, fast and localized accident recognition with reduced latency has also become possible [64]. To further improve accident detection performance and road safety, other studies focused on two-stream convolutional networks [66], the feature fusion framework [67], pedestrian accident detection models [70], a traffic monitoring system based on AI [72], and an intelligent emergency notification system [73] have been investigated. While these have their benefits, deep learning models are huge in terms of data sets, computational power, and the ability to adapt to weather, lighting, and occlusion changes.

Deep learning-based methods have a number of drawbacks, despite their better performance. One significant challenge is the high computational complexity, especially for real-world applications on resource-limited devices. To achieve strong learning results, these models demand high-performance GPUs, high memory capacity, and huge training data sets. However, detection performance can also be influenced by environmental conditions, including lighting conditions, weather changes, occlusion, and camera viewpoint change.

2.2.4 Limitations of Existing Detection Techniques

While various accident detection techniques have been developed, from simple rules-based methods to sophisticated deep learning-based techniques, there are several aspects that

without doubt prevent most of them from being a success in real-world traffic surveillance applications. Current approaches tend to be less successful in more varied traffic situations because of differences in road geometry, traffic volume, climate, and camera perspectives. Traditional and early machine learning-based methods are very feature-driven and rely on handcrafted features and predefined rules that are not sufficient to model complex interactions between vehicles and the dynamics of traffic flow [35].

Labelled datasets are a significant challenge in supervised learning methods due to the rarity of traffic accidents, which makes it difficult and costly to collect large-scale annotated traffic data sets. The datasets used in research have been improved by benchmark datasets like TAD, but the lack of diversity in the datasets remains a problem with model generalization [52], such as the TAD. However, real-time crash prediction on express highways is still challenging because of the rare and arbitrary nature of the accident events on express highways, which causes imbalanced learning problems and less prediction reliability [74].

The high detection accuracy of deep learning methods comes at the same time as other challenges such as computational complexity, scalability, and interpretability. In the development and deployment of deep neural networks, the need for substantial computing power, such as powerful GPUs and ample memory, can pose a challenge for resource-limited environments. However, the challenge to achieve an optimization of detection accuracy and processing efficiency has remained a key research issue in the design of edge computing-based traffic accident recognition systems [64]. Poor illumination, rain, fog, shadows, and occlusion due to environmental changes further compromise the detection performance and impact on the quality of surveillance data. A study of real-time intelligent intersection monitoring systems and a study of feature fusion-based crash detection systems revealed that occlusion and visibility problems can have a great impact on the robustness of the approach [66, 67].

The current approaches also have problems in accurately modelling long-term temporal dependencies in traffic videos, leading to false alarms and missed detections in complex traffic situations. Moreover, most deep learning frameworks are black-box systems, which are not easily interpretable and therefore are not trusted by users in critical traffic safety applications. Although there are explainable AI methods to improve model transparency, their use in accident detection systems remains constrained [48]. A major disadvantage is that none of the integration is done between accident detection, video summarization,

emergency response, and intelligent transportation surveillance systems, which reduces the effectiveness of surveillance systems. These demonstrate the requirement for designing more powerful, enhanced, interpretable, and scalable methods that are able to work in the real road traffic environment.

To conclude, considerable advancement has been made in traffic accident detection, but there is still a need for several challenges to be addressed. There are still some limitations to current approaches, such as their dependence on labelled data, complexity, environmental sensitivity, scalability, and lack of interpretability. Moreover, temporal traffic behaviour modelling and false alarm rate reduction are still big research challenges. The above limitations show that the requirement for advanced accident detection systems that combine spatiotemporal learning, anomaly detection, adaptive capability, interpretability, and computational efficiency in practical traffic surveillance conditions is crucial.

2.3 Video Summarization Techniques

A growing amount of video data is collected by traffic surveillance systems, which has led to a critical need for effective video analysis, storage, and retrieval techniques. Surveillance video includes a considerable amount of redundant and irrelevant information, making it difficult to process manually and computationally expensive to analyse. To overcome these challenges, video summarization has become an important technique to produce short and informative video representations from long video sequences. It can be broadly categorised

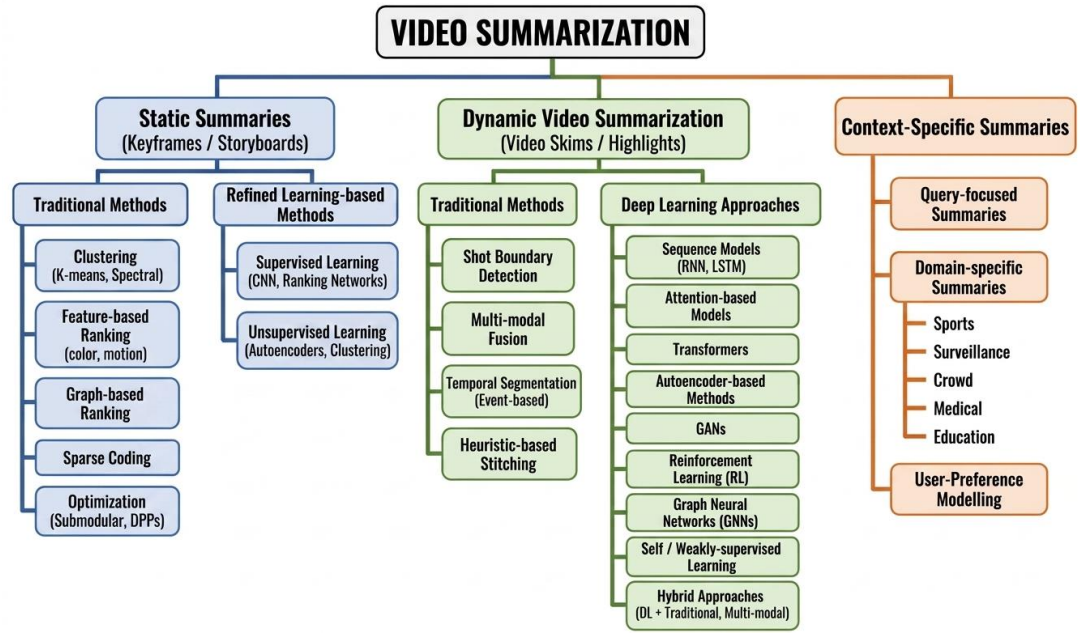


Figure 2.1: Taxonomical Representation of Video Summarization Approaches

into three types as shown in Figure 2.1: static summary, dynamic summary, and context-specific summary. The diagram shows the overview of traditional approaches, learning-based techniques, and advanced deep learning methods used for efficient extraction and generation of concise video summaries.

The purpose of video synopsis techniques is to retrieve the most important information from video data and reduce its length. Summarization, in the area of traffic surveillance, concentrates on identifying significant events like accidents, traffic offences, and unusual activities. Video synopsis can lead to faster analysis times and better storage efficiency in intelligent transportation systems. These methods can also help reduce the amount of data stored and analysed, which can lead to faster analysis times and better storage efficiency in intelligent transportation systems. In the past, different summarization methods for video have been proposed that can be divided into three generic approaches: keyframe-based, event-based, and clustering-based summarization approaches. Each of these techniques has its own pros and cons, depending on the application and the nature of the video data.

2.3.1 Keyframe-Based Summarization

One of the oldest methods and the most widely applied in video summarization is keyframe-based summarization. The main goal of this technique is to minimize video data by selecting a subset of representative frames that can represent the important visual information and overall content of the video sequence. Key frame-based techniques are also used to reduce storage and computation requirements, while contributing to the faster browsing and retrieval of videos, since only significant frames are stored. These methods are simple and efficient; thus, they are widely used in surveillance systems, multimedia retrieval, and video indexing systems.

Several researchers have contributed to the development of summarization methods for video surveillance and analysis that are based on a keyframe-based approach. Kotkar and Sucharita proposed a scalable framework for identifying anomalies using keyframe extraction and applying machine learning algorithms for efficient video surveillance [75]. They found that only if the data is well represented through the selection of a representative frame can it be used for anomaly detection tasks with a significant reduction of redundant data. Kadam et al. proposed a behavioural profiling-oriented adaptive video summarization technique to shift the focus of video summarization from general to personal summary of a

video according to the user's behaviour and preference [76]. Furthermore, Sun et al. proposed a visual framework for video summarization, named VSumVis, to better understand, visualize, and diagnose the video summarization models, which can help increase transparency and interpretability in video summarization systems [77].

2.3.2 Event-Based Summarization

Event-based summarization is an advanced video summarization method that aims at detecting and summarizing meaningful events in video sequences instead of choosing individual representative frames. Event-based summarization aims at semantic understanding, focusing not only on keyframe selection, but also on identifying critical events like accidents, abnormal behaviour, crowd movements, or traffic infractions.

In modern research trends, there have been studies that investigate advanced DL and reinforcement learning to improve video summarization by events. Xu et al. designed a deep reinforcement learning approach for crowd-aware surveillance video summarization, training a surveillance system to identify which video segments are important for generating an informative summary [78]. In a similar way, Liu et al. presented a video summarization approach using reinforcement learning that considers spatiotemporal characteristics of a video by using a 3D spatiotemporal U-Net [79]. Wang et al. also suggested a progressive reinforcement learning approach to gradually remove redundant information and efficiently identify informative video segments over time [80]. They present the new important role of intelligent techniques in the summarization of events in video to be used in recent video surveillance applications.

In traffic surveillance systems, event-based summarization plays a vital role in the detection of segments related to accidents and timely analysis and response. Most of these methods involve anomaly detection algorithms as one of their constituent parts, which can be used to identify events of interest and select the relevant video segments. Event-based summarization allows for more efficient storage management, quicker video analysis, and better decision-making in an intelligent surveillance system by focusing on key moments [35].

There are some benefits to the event-based summarization approach over more traditional summarization. It maintains the temporal signature of an event, keeping the temporal order of actions that result in an incident. This is especially important when considering the context

of an accident and when conducting detailed analysis. It is also more informative, since the summary is based on information that is significant to the event, rather than on individual frames. However, there are challenges associated with event-based summarization. Accurate event detection demands powerful models that can manage various traffic situations and environmental changes. In addition, the efficient integration of the processes of detection and summarization is still a complex challenge, as these processes are often developed independently. To conclude, there are still some limitations to current approaches, such as their dependence on labelled data, complexity, environmental sensitivity, scalability, and lack of interpretability.

2.3.3 Clustering-Based Approaches

The clustering-based approaches are one of the important categories of video summarization techniques because of its ability to reduce video sequence redundancy without losing the most informative content of the videos. They work by clustering frames or segments visually or semantically similar by extracting feature representations and selecting some representative frames or segments from each cluster to produce a short summary. Clustering-based summarization reduces the storage space requirement, computational burden and enables quick retrieval and analysis of videos. These are especially useful in large-scale surveillance systems where multiple cameras are recording the same video content in real time, resulting in redundant data.

There are many different approaches to clustering, including k-means, fuzzy clustering, density-based and hierarchical clustering for video summarisation applications. The effectiveness of these strategies relies heavily on the quality of the extracted spatial and temporal features, similarity measures and methods for optimizing clusters. In [81] Taheri Sarteshnizi et al. presented a pairwise hybrid clustering approach to detecting useful traffic behaviour patterns in large temporal datasets of urban traffic volumes, showing the usefulness of clustering techniques to detect meaningful temporal patterns from large temporal datasets. In same way, Rabbouch et al. designed an unsupervised video summarization method based on cluster analysis for automatic vehicle counting and recognition, and they used the cluster analysis methods to select the representative traffic scenes and eliminate redundant information in video summary [82].

Recently, with the advancement of technology, the clustering-based summarization has been reinforced again, using convolutional neural networks and Spatio-temporal representations to extract high-level semantic features. But there are still a number of challenges, like optimal cluster selection, computational complexity, and processing high dimensional video data, that are still important research issues.

2.3.4 Video Summarization for Traffic Accident Analysis

During the development of ITS, with the increasing use of smart traffic surveillance and the continuous generation of large-scale traffic video data, the video summarization of traffic accidents and traffic incidents has become an important research field. Unlike the traditional approaches, which are mostly focused on minimizing the visual redundancy in video, the goal of traffic accident summarization is to identify the important traffic situations, including traffic accidents, abnormal vehicle movements, heavy traffic, and road events. These techniques aim to generate brief, informative summaries with semantically relevant events for efficient monitoring, forensics, and emergency response.

In recent years, more research has begun to combine deep learning, anomaly detection, and intelligent event analysis into the traffic video synopsis system. Thomas et al. presented a perceptual video summarization framework for road event detection, which included perceptual importance measures to compute the importance of various road events in surveillance videos [136]. Kosambia and Gheewala proposed a video synopsis framework to detect accidents in surveillance videos using deep learning techniques and proved that intelligent summarization can substantially shorten the surveillance video length without losing the information about the accidents [137]. Based on Spatio-temporal rough fuzzy granulation using Z-numbers, Pramanik et al. proposed a traffic surveillance anomaly detection and video summarization method that enhanced the representation of uncertain and complex traffic events for surveillance systems [138].

Summarization techniques focusing on privacy protection and learning have also been of great interest in recent years. Werehti et al. suggested a privacy-preserving road traffic event summarization model for smart cities based on IoT, where the need for secure handling of surveillance data in the context of smart cities is emphasized [139]. Likewise, Saxena and Asghar adopted the YOLOv5 deep learning system for road event-based video summarization to effectively detect and extract significant road events from a continuous

surveillance stream [140]. Although these methods have been proven effective, there are still further research opportunities in the intelligent traffic surveillance system to solve problems such as realistic processing, anomaly understanding, environment adaptability, detection, and summarization.

2.3.5 Challenges in Video Summarization

Despite the progress in video summarization technologies, a number of key research challenges and gaps exist in several aspects of this technology, which limit its effectiveness for smart traffic monitoring applications. One of the biggest problems is the capacity to remove irrelevant information and keep semantically relevant events, such as accidents, traffic misbehaviours, and unusual behaviours of the vehicles. In a continuous surveillance stream, there are many frames that look similar by visual appearance, and it is hard to distinguish meaningful frames from those of repetitive background activities [75, 82]. This results in the loss of important context detail in many current summarization algorithms or the generation of redundant information within the summary.

It's another big challenge to get the balance right between compression and information retention. If the compression is excessive, important event information may be lost to the investigator, and traffic analysis of the event may be hampered if information is not summarized enough. Furthermore, keeping the temporal order fixed is a challenge in many existing and frame-based summarization techniques that tend to lose the order of events in a timeline required for comprehending the accident development and behavioural trends [77, 79].

Another drawback is the lack of real-time implementation of the video summarization frameworks. Many recent methods based on deep learning, reinforcement learning, and anomaly detection for summarization all require significant computational power, processing capacity, and large amounts of annotated data to be trained and tested [78-80]. Such requirements prevent their use in resource-limited edge devices and widespread smart city surveillance networks. In addition, the illumination change, bad weather, shadows, occlusion, camera movements, and high traffic density have a severe impact on the feature extraction and event detection accuracy [136-140].

Another major gap in research is the lack of adaptability and generalization of current summarization models for multi-modal traffic. The methods have many that are highly data-

dependent and don't always work in out-of-sample road conditions or different surveillance situations. Furthermore, existing systems address several aspects of video processing separately, leading to disconnection and high computational complexity in the system architecture for the detection, anomaly analysis, and summarization of accidents. There is still a significant gap in the research of traffic surveillance applications, particularly in the field of multi-agent end-to-end systems that are able to detect and analyse events and then summarize them intelligently.

2.4 Spatio-Temporal Modelling Approaches

Spatial-temporal modelling is a key part required for the current video-based traffic surveillance systems, which allows learning both spatial properties and temporal properties of traffic scenes simultaneously. The fundamental data of traffic video is therefore both motion and visual, with visual data in each frame capturing spatial aspects (such as the appearance of objects, road surfaces, and contextual elements) and motion capturing temporal aspects (such as the behaviour of objects in motion, their interactions, and the patterns of their movements over time). However, comprehensive understanding and effective detection of accidents and anomaly detection require a combination of both spatial and temporal representations.

The previous traffic monitoring methods mainly centred on the extraction of spatial features by image analysis techniques. While they were successful in object detection and scene understanding, they failed to capture temporal dependencies and the evolution of movement across consecutive frames. Thus, purely spatial methods were not often successful in determining collision, unusual driving behaviour, or erratic vehicle behaviour. Likewise, temporal-only techniques were insufficient for using visual context to interpret the scene. This led to the development of the research track of Spatio-temporal modelling, which allowed for modelling all three aspects of appearance, motion, and event progression in the surveillance videos.

The combination of convolutional and sequential learning architectures has greatly enhanced the Spatio-temporal modelling capability. The integration of CNN and recurrent networks, transformer-based models, Autoencoder-based models, and attention mechanisms has significantly enhanced capabilities in traffic analysis, accident detection, anomaly detection, and behaviour understanding tasks [48, 51, 52]. The aforementioned techniques are able to

create powerful feature representations that can learn complex spatial structures and long-term temporal dependencies in traffic conditions.

2.4.1 CNN-Based Models

Spatial feature extraction in video surveillance and traffic analysis systems relies on the primary foundation of CNNs. These convolutional architectures are able to learn the hierarchical visual representations automatically by performing convolutional operations, and the low-level and high-level features like edges, textures, object contours, vehicle shapes, and scene structure can be extracted. CNN models are widely applied in various transportation tasks, such as vehicle detection, traffic monitoring, road object classification, lane analysis, and accident recognition for intelligent transportation systems.

One significant benefit of CNN-based methods is that they are able to derive discriminative semantic features directly from raw image information. In the field of surveillance, advanced architectures like VGGNet, ResNet, InceptionNet, EfficientNet, and YOLO based models have shown great improvement in traffic scene understanding and object detection accuracy [49, 83]. These models can detect complex traffic patterns, vehicle behaviours, and road infrastructure parts under various environmental situations. CNN-based methods have also been used in various applications like traffic flow estimation [41], detection of emergency vehicles [62], traffic violation monitoring [87], and intelligent accident analysis [89].

Recent advancements have further improved the CNN performance by incorporating transfer learning, multi-scale feature extraction, and real-time detection mechanisms. High-speed video object detection frameworks have proven their ability to detect fast and accurately, which can be applied to traffic surveillance applications to decrease the computational latency [88]. Also, CNN-based methods have been useful in classifying accident images [48], detecting road events [50], and intelligent traffic monitoring systems [86].

Although CNN-based models have these advantages, they have limitations in their temporal understanding since they simply process each frame without explicitly modelling sequential dependencies between frames. As a result, they might not properly detect the motion evolution and progression of events in dynamic traffic situations. In the case of distinguishing between regular braking and an actual accident, time-series analysis of several frames may be necessary and is often more helpful in determining than would an individual

frame analysis. This limitation requires temporal modelling architectures to be coupled with CNN-based spatial feature extraction architectures.

2.4.2 LSTM and BiLSTM Models

CNN-based models are limited in time; thus, RNN-based approaches are extensively utilized for sequential analysis of traffic videos, such as long short-term memory and bidirectional long short-term memory. These models are particularly used to process temporal data, as they can easily store a history of data through internal memory representations, thus allowing temporal dependencies and motion dynamics to be modelled over time.

LSTM networks employ the concept of gating and memory cells: input, forget, and output gates regulate the transmission of information in a sequence. This capability enables the network to retain pertinent temporal information while rejecting irrelevant patterns. In traffic surveillance applications, LSTM models can process vehicles' trajectories, motion transitions, sudden behavioural shifts, and changing traffic patterns of accident scenarios and anomaly events. This temporal learning ability is very beneficial in terms of accident detection on video-based surveillance systems.

The BiLSTM models also benefit from their temporal modelling since they take sequences as input and process them both forward and backward. In contrast to traditional LSTM models, which only use the past context, BiLSTM models take into account both past and future context to gain more complete knowledge of the event progression. This two-way analysis is very useful in accident detection tasks that can be used to better classify incidents and recognize anomalies when understanding the context before and subsequent to the event's occurrence.

In recent years, hybrid CNN-LSTM and CNN-BiLSTM models have been shown to achieve better results in Spatio-temporal traffic analysis tasks, where the spatial features of the image are extracted using CNN and the temporal sequence is modelled using LSTM or BiLSTM [51, 53]. In addition, attention-based mechanisms and transformer-like architectures have been used to enhance the learning of long-range temporal dependencies and adaptive feature representation. For traffic anomaly detection, selective attention mechanisms such as multi-head self-attention and hierarchical feature reconstruction have been proven to be effective when they focus on the spatio-temporal patterns that are relevant to the problem [84]. These

methods provide better interpretability, anomaly localization, and temporal consistency in complex traffic environments.

2.4.3 Attention Mechanisms and Transformer-Based Models

Recurrent architectures, despite their success in sequence modelling, tend to lose information when the temporal gap between related events grows large. Attention mechanisms were introduced to address this weakness by allowing a model to weigh the relevance of every frame or feature against every other, instead of relying on a compressed hidden state passed step by step. In traffic surveillance, this ability matters because the visual evidence of an accident: sudden deceleration, trajectory deviation, post-collision stillness, may be separated by dozens of frames.

A recent survey by Sabale and Khedkar [4] catalogues the growing family of transformer variants applied to surveillance anomaly detection and reports their strong accuracy, while also pointing out their heavy computational demands. Building on this direction, Natha et al. [22] fused YOLOv8 detections with a CNN-Transformer pipeline for end-to-end road anomaly recognition, demonstrating that object-level cues and global attention can be combined within a single framework. Attention has also proven useful outside the visual domain; Liang et al. [28] employed multi-head self-attention over 1D convolutional features to detect abnormal highway sounds, showing the mechanism generalises across modalities. Addressing the efficiency concern directly, Chen et al. [84] proposed TA-NET, which uses multi-head local self-attention with adaptive hierarchical feature reconstruction to achieve efficient traffic anomaly detection.

Attention has likewise been embedded within recurrent designs. Natha et al. [103] added an attention layer to a BiLSTM so the network could emphasise frames carrying anomalous motion while suppressing routine traffic, improving both spatial and temporal anomaly detection. Similarly, Hameed et al. [122] paired a vision transformer with a recurrent network, using the transformer to extract rich spatial descriptors and the RNN to track their evolution over time. Across these studies, the trade-off between representational power and inference cost remains central when such models are considered for deployment on live camera feeds.

2.4.4 Temporal Dependency Challenges

Even though many developments in the field of Spatio-temporal modelling techniques have been made, there are still some problems that make it difficult to rely on and scale intelligent traffic surveillance systems. A key challenge is the modelling of complex motion patterns that occur in practical traffic situations. The temporal representation of traffic scenes is extremely complex, as several interacting objects, irregular trajectories, different speeds for the vehicles, occlusion, and abrupt behavioural changes are present.

Another challenge in sequential video analysis is the learning of long-term temporal dependencies. While LSTM and BiLSTM networks are suitable for learning temporal relationships, they could suffer from vanishing gradients, memory constraints, and higher computational complexity in long video sequences. As a result, significant context information surrounding the slowly changing anomalous events might be overlooked in the course of sequence modelling.

Another key drawback with large-scale traffic surveillance systems is computational efficiency. However, taking into account the real-time application of continuous high-resolution video streams in ITS and edge-based environments, significant memory, processing, and training resources are required, which limits the applicability of deep Spatio-temporal architectures [35, 37, 64]. In addition, the different environmental conditions, including changes in illumination, bad weather, shadows, point of view changes, and video quality, have a significant impact on feature extraction and temporal consistency.

The Class imbalance between normal and abnormal traffic events also poses a challenge for model learning. Deep learning models can be heavily biased towards the dominant normal traffic scenario and consequently have a limited ability to detect anomalies and be sensitive to them, due to the fact that accidents and anomalies are relatively rare events. Moreover, most of the current Spatio-temporal models exhibit poor generalization ability in heterogeneous traffic datasets and different surveillance conditions.

To overcome these limitations, efficient, adaptive, and generalized frameworks that can perform robust Spatio-temporal learning under a variety of traffic scenarios need to be developed. Attention, transformer networks, unsupervised representation learning, and hybrid anomaly-driven approaches have emerged in recent years, with potential applications

that offer an exciting outlook for enhancing the temporal dependency modelling and real-time accident detection of intelligent traffic surveillance systems.

2.5 Autoencoder-Based Anomaly Detection

Anomaly detection using Autoencoders is a significant research field of the intelligent traffic surveillance system and can be used to learn the normal traffic behaviour without supervision. Autoencoder-based frameworks, on the other hand, are trained from unlabelled data to reconstruct the input information, requiring no large, annotated datasets for supervised learning. These models are typically trained with normal traffic patterns, and during inference, when there are anomalous events like accidents, sudden vehicle actions, and unusual traffic behaviours, the reconstruction error is a lot higher.

The advantage of the autoencoder-based techniques is to be able to learn complex spatiotemporal features from large-scale surveillance video data without extensive manual labelling. This is of great benefit in the field of accident detection systems where abnormal events are rare and hard to label. Recently, the critical role of deep learning-based anomaly detection systems in traffic accident analysis, surveillance monitoring, and intelligent transportation systems has been proven [51], [54-56]. These strategies offer strong representation learning power for the detection of abnormal traffic events under various conditions.

2.5.1 Convolutional Autoencoders (CAE)

Convolutional Autoencoders (CAE) are well suited DL approach for incident detection in intelligent traffic surveillance applications because they can learn a compact and meaningful spatial representation for high dimensional image and video data. In contrast to the conventional autoencoders, the CAE architectures feature convolutional operations that allow for extracting and reconstructing features, which helps preserve spatial correlations and structural information in traffic scenes efficiently. These features underscore the significant suitability of CAE models for surveillance applications such as accident detection, abnormal vehicle behaviour detection and traffic anomaly detection.

In this case, when applied to traffic surveillance, CAE-based architectures will be trained with a traffic flow of normal traffic to learn the representative distributions of latent features. Anomalous events like collisions, sudden vehicle deviations, or irregular motion patterns, as

well as traffic disruptions, result in much higher errors in the reconstruction during inference when the learned model is not able to reconstruct unseen anomalous patterns. By extracting discriminative visual features from surveillance data with deep convolutional techniques, traffic detection and classification have been tackled with success [49, 50].

Recent studies have also enhanced the capabilities of the CAE-based anomaly detection by introducing temporal learning and an event-driven analysis framework. Using a hybrid CNN-VAE architecture to get better latent representation learning has been effective in addressing vehicular trajectory classification and traffic anomaly detection [51]. Furthermore, to achieve automatic traffic event detection in real-time, temporal convolutional autoencoder networks (TCANs), which can better capture the Spatio-temporal traffic patterns, have been suggested to improve the performance of the traffic anomaly recognition in complicated traffic surveillance scenarios [85].

While the use of CAE-based approaches has been found to be effective, there are several drawbacks. Many of the conventional convolutional autoencoder models are focused on spatial reconstruction and do not sufficiently model long-term temporal dependency between the video sequences. Furthermore, environmental variations, such as illumination change, occlusion, weather effects, and dynamic traffic density, are still important issues in real-world ITS that affect their performance.

2.5.2 Variational Autoencoders (VAE)

Variational Autoencoders (VAE) are the probabilistic version of autoencoder architectures and have been applied to traffic anomaly detection research because they are generatively learned. Unlike conventional autoencoders, which learn deterministic latent representations, VAEs model the latent feature space as a probability distribution, which allows better generalization and robust representation learning in the presence of varying traffic conditions.

VAE-based systems are often used in intelligent traffic surveillance systems to learn the distribution of normal traffic patterns and identify abnormal patterns by looking for abnormal deviations in the latent representation or in the reconstructed patterns. This ability to model in a probabilistic way enhances robustness when dealing with complex traffic situations, environmental changes, and irregular traffic behaviours. Recent studies showed that the hybrids of CNN-VAE can effectively improve vehicular trajectory classification and

traffic anomaly detection by extracting spatial features using CNN and modelling the latent space in terms of probability using VAE [51]. These approaches work well for representation learning and anomaly discrimination in the analysis of surveillance videos.

There are several benefits to the VAE-based models, such as latent space regularization, better generalization of features, and more adaptability to unseen traffic scenarios. These features enable them to be ideal for an unsupervised anomaly detection system where labelled accident data is scarce. Despite these advances, there are several limitations in current VAE-based methods, including degraded reconstruction quality, lack of fine-grained spatial details, and high computational complexity, especially for high-resolution and realistic traffic surveillance scenarios. Furthermore, it is still a challenge to capture the long-term temporal dependency in complex traffic scenes.

2.5.3 Reconstruction-Based Detection Methods

Reconstruction-based methods are one of the most commonly used methods in the area of autoencoder-based anomaly detection for intelligent traffic surveillance systems. The principle behind these methods is that models learned from normal traffic conditions can accurately reconstruct a normal traffic scene, while anomalous events will generate much larger reconstruction errors from learned models because they will deviate from the learned traffic models. The mean squared error (MSE), structural similarity index (SSIM), and feature-level reconstruction loss are typically used for reconstruction error metrics, which are employed to detect abnormal traffic events like accidents, unusual or irregular vehicle movements, and traffic behaviours.

Recently, the combination of reconstruction-based frameworks with deep Spatio-temporal learning architectures has attracted increasing research interest to achieve better performance in anomaly detection in complex surveillance environments. The hybrid model CNN-VAE has shown its ability to classify vehicle trajectories and detect traffic anomalies with improved latent feature learning and reconstruction analysis [51]. Additionally, deep learning accident detection systems have enhanced the detection of anomalies in traffic monitoring videos with adaptive feature extraction and smart event analysis mechanisms [54]. Furthermore, deep ensemble models and anomaly-based traffic monitoring systems have demonstrated good performance in detecting abnormal traffic behaviours in complex traffic and environmental scenarios [55, 56].

Although reconstruction-based methods are effective, there are several important challenges that still remain. If the reconstruction errors are not significantly different in normal and abnormal events, determining whether there is an anomaly in the reconstruction is difficult. Moreover, selecting the optimal thresholds for anomalies remains a challenge to the reliability of anomaly detection and robustness of the systems. In addition, such frameworks are susceptible to the impacts of lighting conditions, climate, shadows, occlusion, and camera noise, thereby reducing the detection accuracy in real-world traffic surveillance applications.

2.6 Comparative Analysis of Existing Methods

After an extensive survey of current traffic surveillance and accident detection methods, it can be seen that there are significant variations in the accuracy of detection, complexity of computation, scalability, and applicability in real-time. The classic approaches to surveillance mainly consist of handcrafted rules (that define the anomalies to monitor), statistical modelling, and predefined thresholds for anomaly detection [38, 40]. While these approaches are comparatively efficient and easy to implement, they tend to have limited adaptability in complex and dynamic traffic environments due to their reliance on manually designed features and static decision rules [43, 45].

Data-driven modelling capabilities for traffic behaviour analysis and accident detection were added using ML-based techniques. Feature extraction and classification algorithms performed better than traditional methods in detecting the targets [48, 50]. These models, however, are still very sensitive to feature engineering and to labelled datasets, which reduces their ability to be generalized to other surveillance scenarios. Moreover, traditional ML approaches are limited in their ability to model temporal relationships in video frames, which is a significant disadvantage for analysing dynamic traffic events.

Recent studies [51, 53, 54] have shown that DL methods can be effectively used to automatically learn spatial and temporal features for traffic anomaly detection. CNN-based architectures, recurrent models, hybrid CNN-LSTM frameworks, and attention-driven networks have proved to be successful in the recognition of accidents and analysis of traffic behaviour, with a high level of accuracy being obtained [66, 67, 84]. Moreover, object detection frameworks like those based on YOLO have contributed to the development of real-time surveillance systems that achieve more accurate detection and localization in less

time [60, 68, 69]. Although these methods are effective, they involve high computational processing speeds, a large amount of data, and a lot of computing power, limiting them to use in resource-limited systems [64].

In recent years, developments in IoT and edge computing have improved the responsiveness and scalability of the intelligent traffic surveillance system, which can process data in a decentralized manner and make decisions locally [35, 36, 37]. They enable real-time monitoring of traffic and minimize communication delays in extensive surveillance systems. Transformer- and attention-based methods [4], [22], [84], [103], [122] outperform plain CNN or LSTM pipelines by capturing long-range temporal dependencies through self-attention. However, their quadratic complexity and high memory demand limit real-time use on edge hardware. Hybrid designs like TA-NET [84] and CNN-Transformer fusion [22] ease this burden, yet performance on continuous multi-camera streams remains unverified.

A comparative analysis of the reviewed methods for accident detection in ITSs, such as fuzzy logic-based model, Bayesian-based model, deep learning-based model, anomaly detection-based model, edge computing-based model, and object detection-based model, is given in Table 2.1. They are compared based on their strategies for detection, their significant contributions, their areas of limitation, and their research gaps. The analysis reveals a consistent trend from the rule-based approach toward complex DL networks and hybrid spatiotemporal approaches to increase the accident detection accuracy.

Table 2.1: Comparative Evaluation of Existing Traffic Accident Detection Techniques

Ref. No.	Accident Detection Approach	Detection Technique/ Model	Key Contribution	Limitations	Research Gap
[43]	Fuzzy Logic-Based Accident Detection	Fuzzy inference system for VANET accident analysis	Improved uncertainty handling in accident detection	Limited scalability and adaptability	Inability to model complex traffic behaviour dynamically
[45]	Fuzzy Bayesian Accident Risk Analysis	Bayesian probabilistic accident estimation	Enhanced probabilistic accident risk prediction	Dependent on predefined assumptions	Limited real-time accident detection capability
[46]	Cooperative Vehicle-Infrastructure Detection	Vehicle-to-Infrastructure (V2I) communication framework	Improved communication-assisted accident recognition	Infrastructure dependency	Scalability issues in large traffic networks
[47]	Traffic Flow-Based Accident Detection	Traffic feature extraction and anomaly analysis	Multi-feature traffic anomaly recognition	Limited temporal modelling	Weak sequential event understanding
[48]	MobileNet + SHAP Accident Severity Prediction	Lightweight CNN with explainable AI	Improved explainability and severity prediction	Requires labelled datasets	Limited temporal dependency analysis
[49]	CNN-Based Traffic Accident Classification	Deep CNN feature map extraction	Automatic spatial feature learning	High computational complexity	Weak long-term temporal learning
[50]	CNN with Mixture of Extreme Learning Machines	Hybrid deep learning classification	Improved accident image classification accuracy	Sensitive to dataset quality	Poor model generalization capability

[51]	Hybrid CNN-VAE Accident Detection	CNN with Variational Autoencoder	Spatial anomaly learning and trajectory analysis	Computationally intensive	Limited real-time deployment
[53]	Video-Based Deep Learning Detection	Sequential video-based accident recognition	Improved accident recognition from video streams	Weak interpretability	Black-box deep learning limitation
[54]	Vision-Based Accident Anticipation	Early collision prediction framework	Accident anticipation before occurrence	High computational requirements	Scalability challenges in real-time systems
[55]	Optimized Neural Anomaly Detection	Optimized neural anomaly learning	Fast convergence and anomaly detection	Threshold sensitivity	Increased false alarm generation
[56]	Deep Ensemble Accident Detection	Ensemble anomaly detection framework	Improved generalization performance	Increased model complexity	High computational overhead
[57]	CCTV-Based Real-Time Detection	Surveillance-assisted emergency response system	Real-time accident monitoring and assistance	Environmental sensitivity	Reduced robustness under varying conditions
[58]	YoFlow Scenario-Based Detection	Scenario-aware roadside accident localization	Improved roadside accident localization	Complex scene dependency	Limited scalability across environments
[59]	Faster R-CNN vs. YOLOv8 Framework	Comparative object detection analysis	Improved evaluation of accident detection accuracy	High resource requirement	Real-time deployment difficulty
[60]	YOLOV8 Real-Time Accident Detection	Real-time object detection model	High-speed surveillance-based accident recognition	Reduced contextual understanding	Weak temporal modelling
[64]	Edge Computing-Based Detection	Edge-assisted intelligent surveillance	Reduced latency and faster inference	Infrastructure dependency	Synchronization and deployment issues

[66]	Two-Stream CNN Accident Detection	Spatial-temporal dual-stream CNN	Improved spatial-temporal feature integration	Computational complexity	Scalability limitations
[67]	Feature Fusion-Based Crash Detection	Multi-feature fusion framework	Balanced detection speed and accuracy	Feature fusion complexity	Generalization challenges
[84]	Attention-Based Traffic Anomaly Detection	Multi-head temporal attention model	Improved temporal feature reconstruction	High training complexity	Real-time optimization challenges
[85]	Temporal Convolutional Autoencoder	Reconstruction-based anomaly detection	Real-time traffic event detection	Limited environmental robustness	Sensitivity to traffic variability

Table 2.2 shows the comparison between existing traffic video summarization methods in intelligent surveillance systems. Keyframe-based, Reinforcement learning-based, clustering-driven, anomaly-aware, object-level, and YOLO-based summarization techniques are also compared. The evaluation reveals substantial progress in summarization using events, in semantic feature extraction, and in traffic video abstraction. However, there are still many existing methods that have high computational cost, limited temporal understanding, low-contextual modelling, scalability problems, and low real-time capability. The above restrictions suggest the necessity of an intelligent and efficient framework to provide traffic video summarization in dynamic surveillance conditions.

The comparative analysis shows that there is no single existing solution that comprehensively and effectively handles real-world traffic surveillance. While deep learning and hybrid frameworks offer promise for enhancing the accuracy of accident detection and video summarization performance, many of these methods still face challenges in terms of high computational complexity, weak temporal modelling, lack of dataset independence, low contextual understanding, and limited generalization capabilities. Likewise, realistic and edge-based systems have quicker processing times but can suffer from environmental fluctuations and scalability restrictions. The challenges highlight the need for an integrated solution that can accurately and efficiently execute traffic surveillance and at the same time have the capability of efficient, robust spatiotemporal anomaly detection and intelligent, event-based video summarization.

Table 2.2: Comparative Analysis of Existing Traffic Video Summarization Approaches

Ref. No.	Traffic Video Summarization Approach	Technique / Model	Key Contribution	Limitations	Identified Research Gap
[23]	YOLO-Based Traffic Video Summarization	YOLO with multi-level masking	Efficient traffic event extraction and surveillance summarization	Increased processing overhead due to masking operations	Limited scalability for dense traffic videos
[75]	Keyframe-Based Traffic Video Summarization	Keyframe extraction with machine learning	Reduced redundancy in surveillance videos	Limited contextual and temporal understanding	Weak event continuity modelling
[76]	Adaptive Video Summarization	Behavioural profiling-based summarization	Personalized and adaptive summarization capability	Increased model complexity	Limited robustness across diverse traffic scenarios
[78]	Event-Based Traffic Video Summarization	Deep reinforcement learning	Improved event-centric surveillance summarization	High computational complexity	Real-time deployment challenges
[79]	Spatio-Temporal Video Summarization	Reinforcement learning with 3D U-Net	Effective spatiotemporal feature learning	Computationally intensive architecture	Scalability limitations
[80]	Reinforcement Learning-Based Summarization	Progressive reinforcement learning	Improved summary optimization and event selection	Long convergence time	High resource dependency
[82]	Clustering-Based Traffic Video Summarization	Unsupervised cluster analysis	Automatic traffic video abstraction and summarization	Limited semantic understanding	Weak anomaly-focused summarization

[101]	Statistical Video Summarization Framework	Statistical taxonomy of summarization models	Comprehensive categorization of summarization approaches	Mainly theoretical evaluation	Limited practical traffic surveillance validation
[124]	Object-Level Video Summarization	Motion autoencoder framework	Unsupervised object-level event summarization	Limited contextual scene understanding	Reduced robustness in complex traffic scenes
[136]	Perceptual Road Event Summarization	Perceptual video summarization framework	Efficient road-event summarization	Limited adaptability under dynamic traffic conditions	Weak anomaly generalization
[137]	Accident-Focused Video Synopsis	Deep learning-based video synopsis	Compact accident-focused video representation	Information loss during synopsis generation	Reduced contextual continuity
[138]	Detection of Traffic Anomaly and Summarization	Spatio-temporal rough fuzzy granulation using Z-numbers	Integrated anomaly detection and video summarization	Computational complexity	Limited real-time processing capability
[139]	Privacy-Preserved Traffic Video Summarization	Privacy-aware IoT summarization framework	Secure road-event summarization	Reduced summarization detail	Trade-off between privacy and summarization accuracy
[140]	YOLOv5-Based Road Event Summarization	YOLOv5 event-driven summarization	Real-time road-event-focused summarization	Object detection dependency	Limited temporal relationship modelling

2.7 Research Gap and Problem Justification

Although tremendous efforts have been made in traffic surveillance, accident detection, anomaly analysis and video summarization, there are still several key research gaps that hinder the effectiveness of the existing ITSs. The lack of a unified approach that combines the accident detection and intelligent video summarization is one of the major limitations. Most of the proposed existing methods process such tasks separately, resulting in a disorganized processing chain, redundant calculations, inefficient surveillance management, and high storage costs [136, 140].

Another important challenge is the heavy dependence on supervised deep learning models requiring large-scale labelled datasets for training [52]. In the practical setting of traffic surveillance, the number of accident and anomaly events is small and cannot be predicted or controlled, making the process of labelling the datasets costly, time-consuming, and hard to generalize to cover heterogeneous traffic scenarios. This results in several existing models being limited in their scalability and lack of robustness to changes in the environment, traffic flow, ambient lighting, occlusions, and weather conditions [56, 74].

Recent CNN-LSTM, CNN-VAE, attention-based, and spatiotemporal deep learning frameworks have significantly advanced sequential feature learning and anomaly detection accuracy [51, 53, 84]. Analogously, the current approaches to traffic video summarization tend to concentrate on redundancy reduction and key frame extraction and fail to consider semantic continuity, anomaly-aware event prioritization, or contextual traffic understanding [75, 78, 136]. Moreover, reinforcement learning, transformer, and attention-based summarization models are often complex and not applicable in real-time settings for large-scale surveillance systems [28, 80, 84, 122].

However, another big drawback is the lack of interpretability and transparency in the DL based surveillance systems. The current accident detection and anomaly analysis methods are all black-box methods, and the interpretation of the decisions is difficult in the application of intelligent transportation systems with safety requirements [48]. Moreover, anomaly detection models based on reconstruction have some difficulties in identifying low-likelihood anomalies as they are easily confused with normal traffic, causing high false alarm rates and poor detection accuracy [55, 56].

It is evident from the above-mentioned limitations that there is a definite need to create an integrated intelligent traffic surveillance framework with the capability for simultaneous anomaly detection and intelligent video summarization that is scalable, robust, and efficient in computation. The proposed research aims to propose an unsupervised spatiotemporal anomaly detection framework using event-oriented video summarization techniques and an autoencoder as a feature learning method. The overall objective of the proposed model is to achieve better detection accuracy, better computational redundancy reduction, better storage space utilization, and better real-time surveillance performance in various traffic situations by combining all four aspects. This complete solution is anticipated to strengthen the advancement of ITS that has strong and scalable capabilities for practical traffic surveillance applications.

Based on the aforementioned problems, there is a need for the development of integrated, scalable and computationally efficient traffic surveillance frameworks capable of simultaneously accomplishing the anomaly detection and intelligent video summarization task. To address these problems, two intelligent traffic surveillance analysis frameworks, CVAE-ADS and Conv-BiLSTM are proposed in the presented research. The proposed frameworks are designed to improve the performance of spatiotemporal anomaly detection, traffic event understanding, and video summarization, and reduce the computational redundancies and maximize the intelligent surveillance capability under various traffic scenarios. These comprehensive systems are anticipated to help build robust and efficient ITS solutions, applicable to real-world surveillance scenarios.

Chapter 3: Theoretical Background and Foundations

3. Theoretical Background and Foundations

The chapter introduces the theoretical background that lays the foundation for the proposed framework for event-based video summarization in traffic surveillance. To understand the design decisions formalized in Chapter 4, it is essential to have a firm understanding of the core concepts, ranging from video data analysis and deep learning architectures to autoencoder models, temporal modelling, clustering, similarity measures, and evaluation metrics. Each section presents the theory, puts it into the context of the overall problem of traffic incident detection and summarization. It provides conceptual foundations for the proposed models: CVAE-ADS and the Conv-BiLSTM autoencoder-based framework.

3.1 Fundamentals of Video Data Analysis

The video data is a series of image frames that contain a series of visual states of a scene at specified points in time, or frames. In a traffic surveillance application, the continuous stream is created using a stationary camera at the intersections in urban areas, arterial roads, expressways, and along highway stretches. To analyse such data requires a knowledge of the spatial meaning of each frame as well as the temporal meaning of the relationship between successive frames. Video streams encode motion, velocity, trajectory, and contextual transitions (e.g., caused by traffic accidents on the road) that are essential for identifying dynamic events, as opposed to still images [95]. The raw video representation may be represented as a 4D array of $T \times H \times W \times C$, where T represents the number of frames, H and W represent the number of pixels in the image's height and width, and C denotes the number of colour channels, typically three for RGB data, used to represent the images. This high-dimensional representation contains a lot of redundancy, as in a normal traffic scene, neighbouring frames are only slightly different, so efficient compression and summarization methods are needed [96].

3.1.1 Characteristics of Surveillance Video

The video produced by surveillance systems has a number of special features that distinguish it from consumer video data in traffic environments. First, recordings are made by stationary cameras, which have a fixed field of view and a fixed number of frames per second (usually 15–30 frames per second). This results in a relatively static background with dynamic

elements (cars, people, and bikes) coming in and out of view. The static background makes the modelling of the background scenes easier, but it also makes the background estimation susceptible to slow variations in illumination, which may damage it [97].

Second, the occurrence of events in surveillance video is highly biased in terms of their timing. The vast majority of recorded footage will consist of normal traffic flow, and abnormal events (collisions, sudden braking, near misses) are unpredictable and rare. Such class imbalance is a structural problem in supervised detection systems since they are optimized for the majority class. It is one of the key reasons for using the unsupervised anomaly detection techniques, in which normal traffic data are used for the training and anomaly detection occurring during test time in both the proposed models [98].

Thirdly, environmental degradation is a common occurrence on surveillance footage. The amount of light that falls on an object can change drastically between day and night, creating a striking contrast in how objects look during the day versus night. In particular, Yang et al. [97] faced this challenge and proposed a depth-aware and domain-adaptive network to ensure the stability of crash detection under different lighting conditions. Other environmental conditions are lens flare, shadowing caused by vehicles and infrastructure, motion blur due to rain, and occlusion (when multiple vehicles are in the same camera frame). These situations add variability, which a good detection system should deal with without deteriorating its performance [99].

Fourth, surveillance footage resolution, compression codec, and frame rates are also different in different deployment scenarios. High-resolution recordings will have more spatial information but will be more demanding of computation, and highly compressed footage might have block artifacts, which will affect feature extraction. This heterogeneity is handled in the proposed framework by normalising all input frames to a fixed resolution of 128 × 128 pixels and normalising the values of each pixel by pre-processing all the frames to be in the range [0, 1] to ensure consistent input statistics, irrespective of the frame originating camera system.

3.1.2 Challenges in Video Processing

The urban-scale processing of surveillance video introduces a number of well-known issues. There's a lot of data produced. Kim et al [96] mentioned that the deployment of a large number of traffic cameras presents a non-trivial infrastructure challenge in the storage and

indexing of the continuous camera data at scale, as well as storing and retrieving the data requires infrastructure development. Within each stream, a camera recording at 720p, 25 frames per second (fps) will result in roughly 1.5 GB of raw data being produced per hour. Thus, a city-wide camera network of hundreds of cameras generates petabytes of video every year, which is far too much to manually review and requires automated video analysis pipelines.

In terms of computation, real-time video processing has very strict latency requirements. The inter-frame time is 40 ms (at 25 frames per second), during which each frame needs to be decoded, pre-processed, and evaluated. This throughput would need to be achieved without compromising detection accuracy through architecturally efficient models. Recent developments in edge computing and systems with 5G connectivity have extended the operating range, allowing the distribution of computation across the network [95, 100]; however, the balance between accuracy and speed is always a design factor.

The fact that surveillance video is redundant adds to the cost of processing. Most of the frames show normal, boring traffic, so performing the same computation on every frame is not a good use of resources. Effective video summarizing can do that and compress by a factor of 70 percent or more while retaining the content significant for post-incident analysis and only those frames that are considered relevant to the event [101, 102].

A further challenge is the need to model temporal context. Accidents are not just single frames: they are sequences of frames and can be identified reliably only by observing the temporal dynamics of the scene. When performed on a frame-by-frame basis without regard to time sequence, such an analysis can miss the pre-impact dynamics, the time of impact, and the post-impact scatter that make up an accident event. This encourages considering sequence-based temporal models, BiLSTM networks, as described in Section 3.5 [103, 104].

3.2 Deep Learning Fundamentals

Deep learning is an umbrella term used for a wide variety of ML methods that utilize multilayer ANN (Artificial neural networks) to extract hierarchical feature representations in data directly from the raw data. It learns task-relevant representations automatically from large training sets, using gradient-based optimization, instead of the need for hand-made features designed by domain experts. It has been successfully used in various applications of computer vision, Natural language processing, and speech processing. Its use for video

traffic surveillance has resulted in significant improvements in detection accuracy and in system robustness [105, 106].

3.2.1 Overview of Deep Learning Architectures

The basic component of a DL model is the artificial neuron, which takes some input, computes a weighted sum of the inputs, which is then passed on to the next layer along with a non-linear activation function. The stacking of neurons in layers and the layers in deep hierarchies enables the model to model very complex, nonlinear mappings between inputs and outputs. Architectural families provide different inductive biases on neuron connectivity, suitable for different data modalities, and enforce such structural constraints.

Convolutional Neural Networks (CNNs) use shared filter kernels to take advantage of the spatially local and translation-invariant nature of image data. This decreases the number of parameters as opposed to fully connected networks while maintaining the spatial structure of the learned features. To process temporal data, architectures with memory, such as Recurrent architectures like LSTM, can be used, since they have a hidden state, which stores information from previous inputs. The most important and central to the present work is the encoder-decoder family, which learns a compact latent representation for the input data and then asks the decoder to regenerate the original input; important to reconstruction-based anomaly detection and latent space-based video summarization [107, 108].

Advanced architectures like GANs (Generative Adversarial Networks) as well as Variational Autoencoders (VAEs) can learn the probability distribution of the training data and therefore generate new data samples similar to the training data. Specifically, here, VAEs can be utilized as a principled probabilistic method for learning smooth, structured latent representations for both reconstruction-based anomaly scoring and clustering in latent space for key-frame selection [109, 110].

3.2.2 Learning Paradigms

Deep learning models can be broadly classified into three learning paradigms: supervised, semi-supervised, and unsupervised learning. The different paradigms assume various forms of labelled training data, and this has various implications for deployment in real-world traffic surveillance.

Supervised learning involves examples of raw inputs and ground truth labels. For example, a supervised accident detector would be fed with a set of images, each with a label indicating whether the image belongs to a normal or accident situation. If there is enough labelled data, supervised models can have a high level of precision. Traffic accidents, however, are rare and labour-intensive to annotate, leading to small, domain-specific, and costly datasets. Models that have been trained on data in one geographic or environmental context are likely not to work well when applied to data from another context [111, 112].

Unsupervised learning does not involve any labelled data. In this paradigm, models are used to model the structure of the training distribution based on the raw observations in the input. Anomaly detection involves building a model that is trained only on normal traffic footage and learns the reconstruction of normal traffic with low error. During test time, frames that are significantly different than the normal distribution to be learned result in more significant reconstruction errors that can be used as anomaly scores. In the process, it has the potential to be scalable because there is an abundance of unlabelled surveillance footage. The detection strategy proposed in this research, based on the reconstruction of the image, fits in this paradigm [109, 113, 114].

3.2.3 Loss Functions and Optimization Techniques

Loss functions and optimization techniques play a crucial role in deep learning systems, affecting a neural network's learning behaviour during training. In deep learning architectures, the primary objective is to minimize the gap between predicted outputs and actual target values by continuously updating the trainable parameters of the network. These can have a profound effect on the convergence, prediction accuracy, and generalization of a DL based Models [105, 114].

- **Loss Functions:**

Mean Squared Error (MSE): This is one of the most widely used loss functions for reconstruction-based approach such as Autoencoders [108]. It evaluates the mean square error of the difference between the given input and the reconstructed output. MSE places more weight on larger reconstruction errors and thus is effective for reconstruction learning and anomaly detection tasks.

Kullback–Leibler (KL) Divergence: For the Variational Autoencoders (VAEs), KL Divergence is employed to regularize the distribution of the latent features, such that it is close to a standard Gaussian distribution [109]. By using KL divergence, the continuity and structure of the latent space can be maintained, thus allowing for efficient feature representation and generation.

Combined VAE Loss Function: The complete loss function of a VAE consists of reconstruction loss and KL divergence regularization loss [113].

- **Optimization Algorithms:**

The process of adjusting the parameters of the model to minimize the loss function is known as optimization in Deep learning. This is usually done by a various algorithm.

Gradient Descent: It is the basic optimization algorithm. It is used to reduce the loss function by taking steps in the opposite direction of the gradient of the loss.

Adaptive Moment Estimation (Adam): This is the widely used optimization technique in deep learning; it has the advantage of computational efficiency and adaptive learning. It adaptively adjusts the learning rate per parameter by keeping an exponentially moving average of the gradients (first moment) and their square (second moment) [115].

AdamW: AdamW is an improved version of Adam that separates weight decay regularization from updating the gradient. In contrast to standard Adam, AdamW applies an L2 penalty directly to the parameters and not in the gradient calculation. It is a more stable method to train networks, particularly in the deep network setting with a large parameter space, and it is a more successful method for obtaining good generalization [108, 118].

Learning Rate (LR) Scheduling:

It is very important to choose the best learning rate for the whole training process. Learning rate scheduling dynamically adjusts the LR in training, which enhances the stability and convergence speed of the training process. In the beginning, a large learning rate is used to learn quickly, and then, a small LR is used in the latter stage to fine-tune parameters. For stabilizing training and performance improvement of the model, various approaches, such as cosine annealing and warm-up scheduling, have been adopted [118].

Regularization and Gradient Control: When there is little surveillance data or the classes are not evenly distributed, regularization and gradient control are important to make sure that the unconstrained optimization is done properly. Regularization techniques are thus used in conjunction with the loss objective:

- **L2 Regularization (Weight Decay):** This is a method that makes sure that the weights do not become significantly large by adding a penalty term in the loss function. This works to decrease overfitting and makes for smoother decision boundaries.
- **Dropout:** in training, randomly drop a small proportion of neurons to ensure that none of the neurons co-adapt, thus strengthening the model's robustness and generalization ability.
- **Batch Normalization:** It normalizes the activations of intermediate layers to have a mean of 0 and variance 1 for each mini-batch. This makes the gradient flow more stable, helps with convergence, and makes it possible to use higher learning rates for training.
- **Gradient Clipping:** This is a method to prevent exploding gradients in some architectures, such as LSTM and BiLSTM networks for long temporal sequences.

As a whole, these loss functions, optimization algorithms, and regularization techniques play a pivotal role in achieving stable convergence, efficient feature learning, and improved generalization. Their incorporation is pivotal in boosting the resilience and performance of DL models for computer vision, incident detection, and smart surveillance systems.

3.3 Convolutional Neural Networks (CNNs)

The most popular family of architectures to process visual data is Convolutional Neural Networks. They can learn hierarchical spatial representations from raw pixel data inputs, which has made CNNs the standard backbones in object detection, semantic segmentation, and video analysis. In the proposed framework, the CNNs are used as the spatial feature extraction component of both the CVAE and Conv-BiLSTM Autoencoder models, which compactly represent the visual information in each frame before temporal modelling or latent space inference [105, 106].

3.3.1 Architecture and Components

A typical CNN consists of alternating layers of convolutional layers, nonlinear activation functions, and pooling layers, and possibly fully connected layers that generate a final prediction.

A convolutional layer is composed of a set of learnable filter kernels, each of size $k \times k \times C_{in}$, and slides over the input feature map, computing dot products between the input patch at each spatial location and the different filters. The output of this operation is a two-dimensional activation map for each filter, which is then stacked together into a three-dimensional output tensor with shape $H_{out} \times W_{out} \times C_{out}$, with C_{out} being the number of filters. Since the weights are shared for all the spatial positions, the number of trainable parameters per position in the convolutional layer is also independent of the spatial size of the input, and therefore, the model can be trained efficiently even for high-resolution inputs [105].

The non-linearity is introduced by activation functions, necessary to enable the network to approximate complex mappings. Rectified Linear Unit (ReLU) is the widely used nonlinearity function in modern CNNs, defined as $f(x) = \max(0, x)$. ReLU has the effect of suppressing all negative activations and keeping all positive ones unchanged, is cheap to compute, and avoids the vanishing gradient problem that can occur with sigmoid and hyperbolic tangent activations in deep networks, thus speeding up convergence in training [115].

Pooling layers spatially downsample by combining the activations in local regions. The most popular one is max pooling, which keeps the max activation value in each non-overlapping pooling window and removes fine spatial details, thus achieving some translation invariance and retaining the strongest activation values. The encoder branch of both proposed autoencoder architectures uses max pooling, which decreases spatial resolution but increases the semantic level of the representations learned.

On the other side, in the decoder part, the transposed convolutional layers reverse the downsampling process by upsampling feature maps back to the original input size. This allows them to reconstruct a complete-resolution frame from a small latent code, as in both models, an anomaly detection approach is based on reconstruction.

3.3.2 Feature Learning Mechanism

One of the most important characteristics of deep CNNs is hierarchical feature learning. The first few convolutional layers capture local features like oriented gradients, edges, and colour changes. These are then used to generate intermediate layers, which represent mid-level information like texture, boundaries of objects, parts of vehicles, etc. The deeper layers are used for the high-level semantic concepts, e.g., vehicle configurations, lane structures, and inter-vehicle spatial relations. This hierarchical representation is especially useful in traffic surveillance applications, as the difference between normal traffic scenarios and accident ones can be detected at the level of scene semantics (e.g, relationships of spatial and contextual relationships between vehicles) [98, 108] and not at the pixel level.

CNN layers used in a time-distributed fashion, i.e., by applying the same spatial CNN operations independently to each frame of a temporal sequence, are efficient spatial feature extractors. The features extracted can then be input to a temporal modelling component for sequence learning. The Conv-BiLSTM Autoencoder design consists of time-distributed convolutional layers, which learn the spatial features from each frame, followed by stacked BiLSTM layers that learn the temporal dependencies across the sequence.

3.4 Autoencoder Models

Autoencoders are a family of neural networks whose overall goal is to reconstruct the input signal, using a low-dimensional bottleneck representation. The network is divided into two parts: an encoder that transforms the input to a compact latent code and a decoder that transforms back to the original input space to create a reconstruction. The model is forced to learn a compact representation that carries a lot of information by discarding redundant or irrelevant details from the bottleneck. Autoencoders trained only on normal data are able to reconstruct normal patterns with low error, while giving high reconstruction error to anomalous data inputs that fall outside of the learned normal distribution. This is a property that renders autoencoders natural choices in the context of unsupervised anomaly detection in surveillance environments [116-118].

3.4.1 Convolutional Autoencoder (CAE)

In the basic autoencoder, the fully connected layers of the original autoencoder are composed of a sequence of layers. An encoder consists of convolutional layers, and a decoder contains

transposed convolutional layers, resulting in a Convolutional Autoencoder (CAE). Unlike the fully connected autoencoders, which would have too many parameters to support high-resolution inputs, this modification maintains the spatial structure of the image data throughout their encoding and decoding process, making CAEs much more suitable for image and video inputs. The encoder of a CAE is a sequence of convolutional and max-pooling layers, which extract the hierarchical spatial features and gradually downsample the spatial dimension. The compact feature map at the bottleneck is the latent representation. The decoder then uses transposed convolutional layers to gradually upsample this representation into the original input resolution to get the reconstructed frame.

Figure 3.1 shows the architecture of the Convolutional Autoencoder model used in this research. Wang and Yang [118] showed the feasibility of using a CNN-Recurrent-based Convolutional Recurrent Autoencoder for video anomaly detection, which jointly models the spatial and temporal dimensions of videos. The standard convolutional autoencoder configuration is used as the Baseline CAE for this research, which is the main CAE model for comparisons with the proposed models.

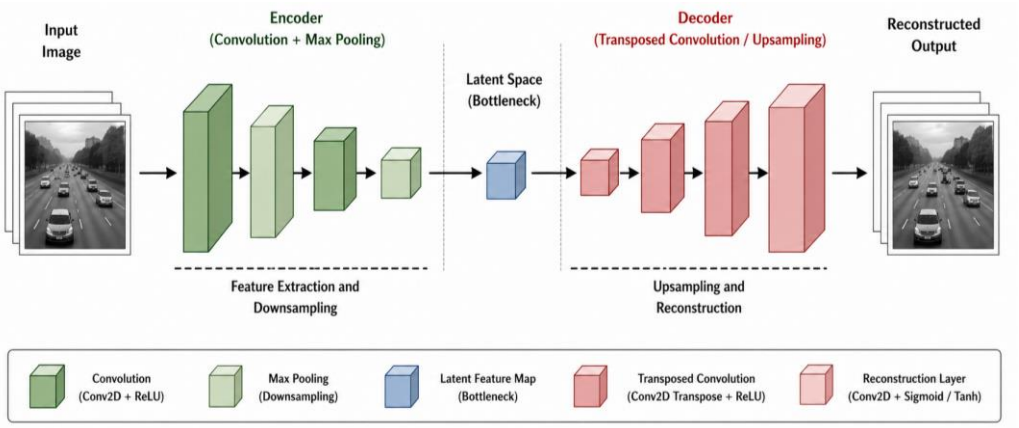


Figure 3.1: Convolutional Autoencoder model architecture

3.4.2 Variational Autoencoder

The Variational Autoencoder is a variant of the Autoencoder in which the latent space is replaced by a probabilistic structure, as shown in Figure 3.2. The VAE is different than a regular autoencoder because when the input is fed into the encoder network, the VAE produces two vectors representing the mean and variance of a probability distribution in the latent space. This is a way of mapping the input to an area of potential representations, not one fixed point. Then, a point in the distribution is sampled by applying the reparameterization method, where the latent vector is found by scaling and shifting the

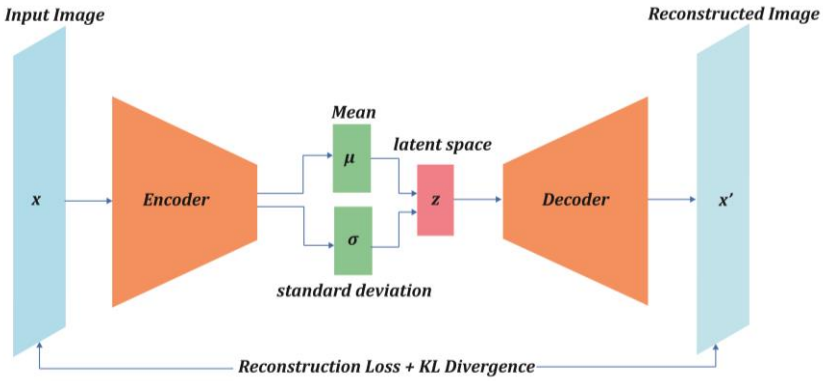


Figure 3.2: Variational Autoencoder model architecture

independently sampled Gaussian random noise with the learned mean and variance. This reformulation allows gradients to pass through the learnable parameters during backpropagation without any restrictions.

The latent vector is then passed to the decoder, which is similar to the encoder but works in reverse, slowly increasing the dimensionality of the compact vector back to the dimensionality of the original input to give the reconstructed output [109]. Optimizing two conflicting goals At the same time is part of training the VAE. The KL divergence loss acts as a regularizer, which ensures the learned latent distributions do not deviate too far from a typical Gaussian prior distribution, and the reconstruction loss ensures that the decoded output is close to the input. This regularization is crucial because it will force the model to structure the latent space in a smooth, continuous, and systematic manner instead of simply memorizing the training data. The latent space is shaped into a useful and navigable representation by striking a balance between these two losses.

Once the training is completed, the VAE can be used as a generative model to generate new samples of data directly from the prior distribution and pass them through the decoder. For nearby latent points, the outputs will be semantically similar. The continuity property is directly used in the video summarization part of the CVAE-ADS model to cluster visually similar frames of the accidents in the latent space to select keyframes using K-Means clustering [114, 121]. Taking a similar approach as the above, Yu and Huang [116] used an encoder-decoder model for anomaly detection of driving trajectories in a spatiotemporal context and again found that the reconstruction error thresholding is generally an anomaly criterion. To validate the theoretical basis of CVAE-ADS, Zhang et al. [117] further extended the VESC model, including variational constraints to enhance the ability of anomaly discrimination.

An and Cho [109] gave a theoretical argument for using the reconstruction probability of a VAE as an anomaly score, which provides the guarantee that inputs with low reconstruction probability are reliably anomalous. Nguyen et al. [113] performed a comparative study, and the result shows that the VAEs-based reconstruction scores have always been better than deterministic autoencoders for anomaly detection tasks. This was later proved to be an effective method by Iqbal and Qureshi [114] on a variety of anomaly detection benchmarks. This was later extended to traffic video specifically by Aslam and Kolekar [121], who introduced an attention-augmented VAE for traffic anomaly detection, which is able to localize the anomaly better.

3.5 Temporal Modelling Techniques

Spatial models, like CNNs, have proven to effectively model the visual content of each image, but they work on each image individually and thus cannot completely model temporal evolution for events in video. Traffic accidents are temporal events in nature: recognising an accident requires following the development of the traffic scene in time: the approach dynamics before the impact, the impact itself, and the scatter of the participants after the crash. Temporal modelling techniques are therefore an important complement to spatial feature extraction in a video surveillance system [103, 107, 122].

3.5.1 Recurrent Neural Networks

RNNs are a family of networks that are built to process sequences of data by keeping an internal hidden state h_t that is passed forward in time and depends on the current input x_t and the previous hidden state h_{t-1}:

$$h_t = f(W_h h_{t-1} + W_x x_t + b) \tag{3.1}$$

Here, W_h, W_x and b are learnable weight matrices and bias vectors, respectively, and f is a non-linear activation function. The hidden state is a vector of accumulated information about the history of the input sequence, which allows the network to make context-dependent predictions. This gives an RNN the ability, in theory to model dependencies between events that occur many time steps apart.

In reality, standard RNNs are known to suffer the vanishing gradient problem: when training, the gradients are propagated backwards over many time steps, and the gradient is multiplied

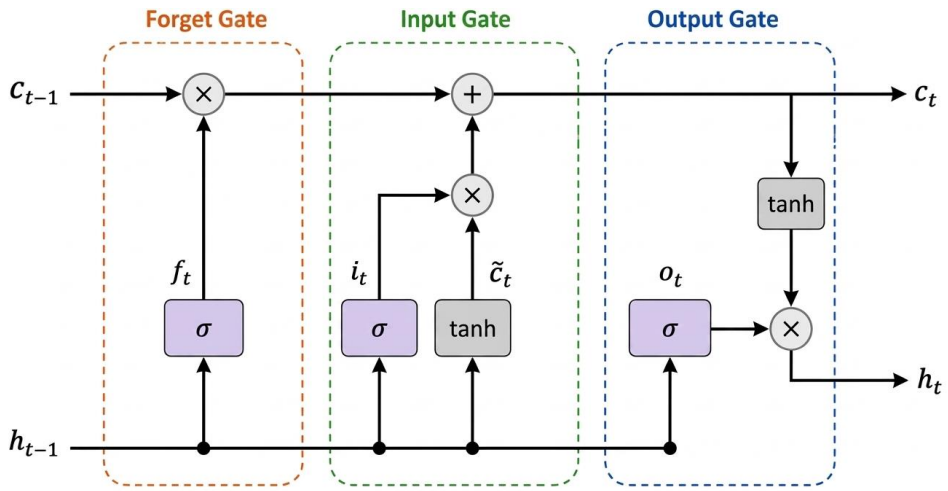


Figure 3.3: Internal Architecture of an LSTM Cell

in each step, making it impossible to learn dependencies of more than a few time steps. This is particularly problematic in the case of accident detection, which may be many frames in advance of the actual visual perception of the accident. The LSTM network was created with this limitation in mind [104].

3.5.2 LSTM and BiLSTM Networks

LSTM network uses a gated memory cell to selectively retain or discard information in long sequences, thus alleviating the vanishing gradient problem. LSTM has two kinds of state vectors: one is c_t, which is used to store long-term information, and the other is h_t, which is used to generate short-term output. As seen in Figure 3.3, there are three different gating mechanisms, which are the forget gate, input gate, and output gate, to control the flow of the information into, through, and out of the cell, respectively. Forget gate: Decides whether to keep or not from the previous cell state.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3.2}$$

Input gate: Supports adding new information to the cell state

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3.3}$$

Output gate: Used to control the information exposed by the hidden state.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3.4}$$

The internal states are updated in the following manner:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \text{ and} \tag{3.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.6}$$

The gating mechanism allows LSTM to selectively remember useful long-range context and to forget unimportant details, which makes it very suitable to model the multi-frame temporal structure of traffic events. Ashraf et al. [120] showed that using an LSTM Autoencoder to learn the normal patterns in ITS data is an effective approach to detecting anomalies, thereby validating the use of LSTM-based reconstruction in traffic anomaly detection. Natha et al. [103] also showed that a deep BiLSTM attention model gives good spatial and temporal incident detection performance for surveillance applications. The BiLSTM is an enhancement of the LSTM, which can process the input sequence in both temporal directions, forward and backward, simultaneously. There are two independent LSTM networks: forward network (t =1, ..., T) and backward network (t = T, ..., 1). The hidden states of both passes are then joined together at every time step to make up the representation, as shown in Figure 3.4:

$$h_t = \left[h_t^{(\text{forward})}; h_t^{(\text{backward})}\right] \tag{3.7}$$

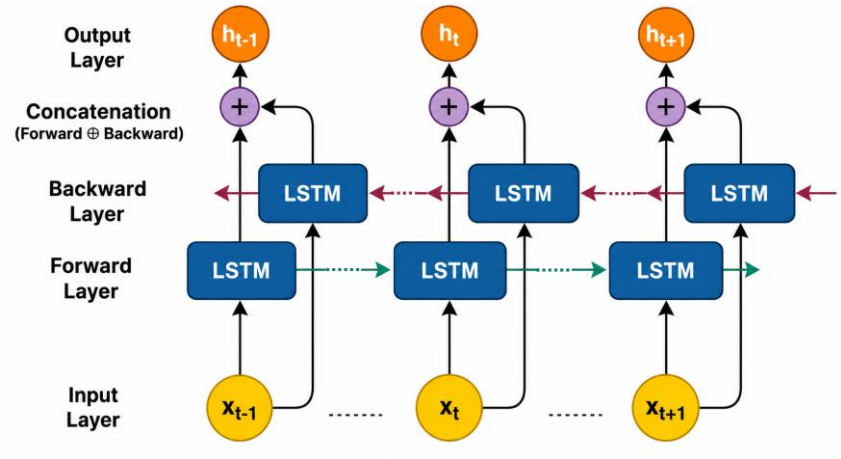


Figure 3.4: Bidirectional LSTM (BiLSTM) Architecture

This bi-directional structure gives the representation of each frame context from the past and the future in the sequence window. For accident detection, it could be that the model can consider the dynamics of the vehicles before and during the accident to describe the frame, significantly enhancing the accuracy of the accident detection compared with the unidirectional LSTM that only observes past information. In the field of summarization, Zhang et al. [107] showed this advantage in the video summarization domain where LSTM-based sequence models can better capture temporal relationships between frames than frame-independent approaches. This was further extended by Lin et al. [104] with hierarchical

LSTM networks, which bring together multi-scale temporal modelling and attention mechanisms for video summarization.

3.6 Clustering and Similarity Measures

The goal of video summarization is to compress a bunch of frames into a smaller group of frames that reflects the significant information of the scene. In the case of accident detection, this means picking a few representative images from a series of detected unusual events. This can be considered as a grouping problem, in which frames with similar visual and semantic contents are grouped together and a representative frame is selected from each group.

This is done using clustering and similarity measurement techniques. Clustering is used to cluster frames according to the similarity of the features and similarity measures to prevent redundant frames from being selected for the final summary. Of the many methods, K-Means clustering is a popular technique for clustering representations of features, while the Structural Similarity Index (SSIM) is a suitable method for determining visually similar frames [125-127].

In this work, such techniques are integrated into the video summarization pipeline to guarantee compactness and informativeness of the generated video summary. Chapter 4 shows a detailed implementation of the summarization process.

3.6.1 K-Means Clustering

K-Means is an unsupervised method that is used for dividing the n data points into K different clusters according to the similarity of features [124]. The algorithm clusters a set of data given in a d-dimensional feature space in a way that minimizes the distance between the data points and the centroid of the clusters. It's an iterative process in which there are two steps: 1) Associate each data point with the closest centroid, and 2) update each centroid to the average of all the points it is associated with. This procedure is repeated until convergence is reached.

Clustering is usually performed in the context of summarising video on feature representations of frames as opposed to raw pixel data [126, 127]. This helps to ‘group’ similar events or visual patterns more effectively. A representative frame is then chosen for

each cluster, typically according to how close it is to the centroid of that cluster, so as to best represent its characteristics. The other metric used was Structural Similarity Index (SSIM).

3.6.2 Structural Similarity Index (SSIM)

Once representative frames are found, one needs to remove redundancy so that one doesn't get visually similar frames in the final summary. This is accomplished via similarity measures, which measure the similarity between two images. A perceptual index that is based on contrast structure and luminance information to measure the similarity of an image [125]. SSIM is aligned with human visual perception as opposed to pixel-based error metrics. Candidates are evaluated with SSIM in comparison with the already selected frames. Once the similarity is above a certain threshold, the frame is assumed to be redundant and rejected. This makes sure that only informative frames are included in the final summary, and they should not be repeated [123].

However, due to the SSIM's immunity to illumination and contrast, it is especially appropriate for surveillance videos taken under varying environments. Consequently, it can enhance the diversity and quality of the video summary generated. Through SSIM, every candidate frame is compared with the selected frames. Once the similarity is above a certain threshold, the frame is assumed to be redundant and rejected. This makes sure that only informative frames are included in the final summary, and they should not be repeated. This is due to the fact that SSIM is insensitive to changes in illumination and contrast, rendering it more suitable for surveillance videos recorded under varying environmental conditions [123]. Consequently, it can enhance the diversity and quality of the video summary generated.

3.7 Evaluation Metrics

It is necessary to have a comprehensive assessment framework to determine the effectiveness of the proposed models and to facilitate a fair comparison with existing models. As the framework covers two tasks, accident detection and video summarization, different evaluation criteria are used for each task, consistent with the goals of the respective tasks.

3.7.1 Accident Detection Metrics

The incident detection problem is treated as a binary classification problem: either normal (negative) or accident (positive) [123, 128]. As such, standard classification metrics are applied.

Accuracy is a measure of the overall correct prediction:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{3.8}$$

However, in a scenario of surveillance data, because of class imbalance, accuracy is not enough [128]. Hence, other measures are taken into account.

Precision measures how accurate positive predictions are:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{3.9}$$

Recall (Sensitivity) is the percentage of actual accident frames that have been detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3.10}$$

The harmonic mean between precision and recall is called the F1-Score, which is a balanced measure of performance:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \tag{3.11}$$

The AUC-ROC is a measure of discrimination that is independent of a threshold setting. It is a measure of the probability for the model to give a higher anomaly score for accident frames than normal frames. Scores nearer 1 are signs of better performance. [128]

The Equal Error Rate (EER) defines an operating point of the system where the False Acceptance Rate and False Rejection Rate are identical. An EER close to 0 represents a better balance between false alarm and miss detection rate. These measures collectively make up a comprehensive and diverse assessment of the effectiveness of accident detection.

Also, to gain a deeper insight into classification performance, analysis of the confusion matrix is used to present a detailed understanding of the number of TPs, TNs, FPs, and FNs.

These metrics, combined with each other, allow a detailed analysis of the detection performance in different traffic and dataset characteristics.

3.7.2 Video Summarization Metrics

Similarly, video summarization evaluation should be based on compression efficiency, the visual quality of the summary, and content diversity [124-126].

The Reduction Rate is a measure of the compression rate:

$$\text{Reduction Rate (\%)} = \left(1 - \frac{\text{N}_{\text{summary}}}{\text{N}_{\text{original}}}\right) \times 100 \tag{3.12}$$

The greater the value, the more the length of the video will be reduced, but with a higher level of information retained.

Peak Signal to Noise Ratio is used to measure the quality of the reconstruction.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{{MAX_I}^2}{\text{MSE}}\right) \tag{3.13}$$

Higher PSNR values correspond to a better fidelity between the original and reconstructed frames.

The Diversity Rate is the rate of uniqueness of the selected keyframes:

$$\text{Diversity Rate} = 1 - \text{mean}(\text{SSIM}(\text{f}_\text{i}, \text{f}_\text{j})),\ \text{i} \neq \text{j} \tag{3.14}$$

A high Diversity score indicates that the summary is not repetitive or redundant across different aspects of the event [125-127].

The quality of the clustering in the latent space is measured using the silhouette score [124-126]. Clusters that are close to 1 are well-separated and compact, making them suitable for effective keyframe selection.

$$\text{s(i)} = \frac{\text{b(i)} - \text{a(i)}}{\max(\text{a(i)}, \text{b(i)})} \tag{3.15}$$

where a(i) represent the intra-cluster distance and b(i) is the nearest-cluster distance. Silhouette scores are clearer and more compact when higher.

Overall, these summarization metrics make sure that the generated summaries are shorter, informative, diverse, and visually consistent. The evaluation is holistic because of the inclusion of clustering quality, redundancy reduction/reconstruction fidelity.

Chapter 4: Proposed Model and System Architecture

4. Proposed Model and System Architecture

In this chapter, the proposed methodology for an event-based video summarization system for a traffic surveillance application is presented. It relies on two deep learning architectures, referred to as Convolutional Variational Autoencoder for Accident Detection and Summarization (CVAE-ADS) and Conv-BiLSTM Autoencoder, both of which follow a reconstruction-based unsupervised learning paradigm and are able to detect anomalous events in continuous video streams without relying on annotated training data. The design, architecture, and operational logic of each component of the proposed framework are thoroughly described in the chapter.

The chapter motivates the proposed framework, provides a conceptual overview of the end-to-end system workflow, an architecture description of both the proposed models, the accident detection framework, the video summarization pipeline, and the full algorithmic workflow of the proposed system. All of these elements form a full technical foundation of the proposed system that makes it possible to detect accidents accurately and to create video summaries of raw traffic surveillance footage that are both concise and informative.

4.1 Motivation for the Proposed Framework

One of the challenges of the present time is the huge amount of traffic surveillance footage produced by the modern smart city infrastructure and how to extract useful and actionable information promptly and accurately from the continuous streams of footage [106]. Traffic monitoring systems are typically manual, relying on pre-labelled data sets, and are thus resource-intensive and impractical at scale [112]. This constraint drives the need for an automated and intelligent framework for efficient and accurate identification of accident events and creation of video summaries with minimal human annotation [102].

The key idea behind the proposed framework is that two tasks, typically considered as separate problems in the literature, namely accident detection and video summarization, are combined into an integrated framework. Combined in a single pipeline, both can lead to more efficient operations and summaries semantically meaningful and event-driven, instead of sampled at random [136, 138].

Second, deep learning architectures, especially hybrid models that integrate both CNNs and RNNs, have proven to possess promising modelling capability for the spatial structure and temporal dynamics of video data [118]. This development encourages the investigation of two complementary architectures: CVAE-ADS for learning compact embeddings of traffic scenes and the Conv-BiLSTM Autoencoder for explicitly modelling temporal dependencies between video frames. Both models are based on the reconstruction-based paradigm for anomaly detection, where a model of normal behaviour is learned, and anomalies are detected as deviations from this normal behaviour.

The proposed framework is also motivated by practical deployment constraints. The traffic surveillance system needs to be efficient in realistic and edge computing environments, where timely detection is critical for emergency response and traffic management [77]. The goal of the framework is to create lightweight unsupervised models that do not rely on annotated data and that are efficient, flexible, and can be used in a wide variety of surveillance contexts, as well as to assist traffic decision-making in a timely fashion [98].

4.2 System Overview and Workflow

The proposed system aims at automatically extracting accident event information from a continuous stream of traffic surveillance videos and creating small video summaries that contain the key information of the accident events. The entire process is shown in Figure 4.1 and can be broken down into four stages: frame extraction, anomaly detection using autoencoder reconstruction, threshold-based anomaly classification, and event-driven video summarization.

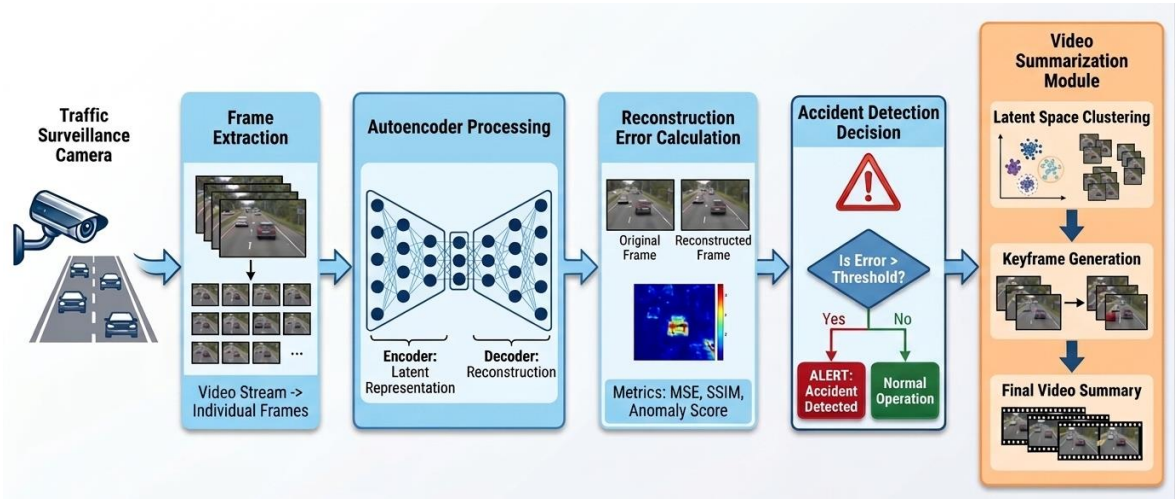


Figure 4.1: Overview of the proposed Autoencoder-based traffic accident model.

- Stage 1- Frame Extraction: The input video stream is first split into frames. The pre-processing of each frame involves resizing and normalizing it to adapt it for deep learning-based processing and minimize the computational burden. This frame-level representation is the fundamental level of analysis used throughout the pipeline.
- Stage 2- Reconstruction-Based Anomaly Detection: These pre-processed frames are fed to one of the proposed two autoencoder models, CVAE-ADS and Conv-BiLSTM Autoencoder. Both models only learn from regular traffic scenes and develop regular traffic pattern restoration. At inference time, the model tries to reconstruct each frame of the input. High reconstruction error occurs for frames that contain abnormal events (e.g., accidents), and high fidelity of reconstruction for frames with normal events. A reconstruction metric is the MSE because it is easy to compute and sensitive to structural deviations.
- Stage 3- Threshold-Based Classification: The computed reconstruction error of each frame is passed to a threshold mechanism. When the error of a frame is above a certain threshold, it will be considered an anomaly, and the flag will be set to accident related. If the frames are below the threshold, they are classified as normal and skipped. The decision threshold is set empirically according to the distribution of the reconstruction errors found in the training phase on normal data.
- Stage 4- Event-Driven Video Summarization: The flagged anomalous frames are sent to the video summarization module. These latent representations of the frames are clustered with K-Means clustering, which will group the frames that have similar visual and semantic characteristics into different clusters. Then, a representative keyframe of each cluster is chosen, usually the keyframe nearest the cluster centroid, which best represents the dominant features of that event segment. Finally, redundant keyframes are removed with the Structural Similarity Index (SSIM) so as to keep only the keyframes that are different in appearance. The final set of keyframes are then organized chronologically to form a summary video that is short and informative, and which shows the evolution of the accident events detected without any repetition.

This complete end-to-end pipeline allows the system to generate structured, event-centred summaries from the raw and unstructured surveillance data, helping traffic monitoring authorities review, make decisions, and archive with efficiency.

4.3 Baseline Models for Accident Detection

Two basic autoencoder models are used as a baseline to assess the performance of the proposed accident detection approach, namely Convolutional Autoencoder (CAE) and Variational Autoencoder (VAE). Both models are based on the encoder–decoder architecture and training them with only normal traffic data can enable them to learn normal traffic patterns. The reason for choosing these models is their powerful ability to detect anomalies that are based on reconstruction, i.e., the reconstruction error from a learned normal pattern is a good indicator of an anomaly, such as accidents.

The baseline models are important for two reasons: (i) the improvements obtained by the proposed method can be compared to them; (ii) they allow for comparison of the deterministic and probabilistic latent representations in the traffic surveillance context.

4.3.1 CAE Baseline

In order to learn the spatial property of a traffic surveillance frame, the Convolutional Autoencoder (CAE) is developed. Contrary to the traditional autoencoders with fully connected layers, the CAE utilizes the convolutional operations, which can well maintain the spatial locality and the hierarchical feature representation. This renders it especially appropriate for pictorial information, such as the traffic scene, where spatial relations (vehicle locations, road layout, movement cues, etc.) are essential.

The CAE architecture consists of two main parts: an encoder and a decoder. The CAE consists of an encoder part consisting of a sequence of convolutional layers (with 32, 64, 128 filters) and max pooling to extract the hierarchical features of the space and compress the input into a latent representation. Convolutional layers learn more abstract features, and pooling layers decrease the spatial dimensions and preserve key features. In this work, RGB frames of the shape (128, 128, 3) are passed through several convolutional blocks, each having a greater feature depth so that the model can learn rich visual representations of normal traffic patterns.

The latent space produced by the CAE is deterministic. The latent vector generated by the encoder is the same for the same input frame. This deterministic mapping can lead to consistent reconstruction outputs and stable reconstruction errors, beneficial for anomaly

detection tasks. The model has only learned to reconstruct normal patterns with high fidelity, as it has only been trained on normal data. But if an abnormal frame (such as an accident) is detected, the model is not able to reconstruct it accurately, and the reconstruction error will be larger.

The decoder is symmetrically built using dense, upsampling, and convolutional layers, and a sigmoid output for pixel-level reconstruction. The CAE's training goal is to reduce the reconstruction error between the original image and the reconstructed image based on pixels, usually by using Mean Squared Error (MSE). This helps the model to learn a compact representation of normal traffic scenes whilst being sensitive to deviations.

The CAE is selected as a baseline because it has high spatial feature extraction ability and is widely used in image anomaly detection tasks. It offers a deterministic reconstruction framework, which is an ideal reference model to determine the effectiveness of spatial information alone in accident detection. Further, it allows exploring if using more sophisticated latent representations (as in VAE) leads to any measurable improvements.

4.3.2 VAE Baseline

VAE is an extension of the standard autoencoder, where the latent space is also given a probabilistic formulation. The VAE does not encode a particular input frame into one single deterministic vector, but into a distribution (usually taken to be Gaussian) representing the latent space. The design enables the model to accommodate the natural variations (illumination changes, traffic density, and environmental settings) and uncertainty (inherent variability) in normal traffic scenes, allowing the model to be more robust.

The input frame is of size 128x128x3, which is flattened and passed to a stack of fully connected (dense) layers in the encoder (as depicted in the architecture). In this case, the encoder outputs two vectors rather than a single latent vector: the mean (μ) and variance (σ^2) of the latent distribution. A latent distribution is then sampled using the reparameterization approach while retaining differentiability and backpropagation compatibility of the sampling operation.

The sampled latent vector is then the bottleneck representation, which is fed to the decoder, which is composed of fully connected layers that gradually reconstruct the input frame. The

decoder aims to learn the mapping from the latent space back to the input space, that is, the original image space.

To optimize the VAE, a composite loss function consisting of two objectives is used. The first is called the reconstruction loss and is an indication of how much the input has been regenerated. The second is KL divergence, which is used to force the learnt latent distribution to be a normal distribution. Together with this dual objective, this means that similar inputs are mapped to neighbouring regions in a structured and contiguous latent space.

This probabilistic modelling results in an organization of normal traffic patterns within the latent space, which is meaningful. At testing time (inference), frames that do not follow these learned frames (e.g., accident frames) are less likely to be well-reconstructed, and these frames may be associated with higher reconstruction errors or lower likelihood scores. The departure of these deviations is a good indicator of abnormal events.

The VAE is used as a probabilistic baseline to investigate the contribution of latent distribution modelling in anomaly detection. The VAE extends the CAE to include uncertainty modelling and latent space regularization, as opposed to the deterministic encoding and spatial feature-learning that are present in the CAE. This can be used to test the effect of capturing variability in normal traffic on the detection performance. The comparison between CAE and VAE, therefore, gives insightful information about the effectiveness of the deterministic and probabilistic representations in unsupervised accident detection.

4.4 Proposed CVAE-ADS Framework

The Proposed Convolutional Variational Autoencoder for Accident Detection and Summarization (CVAE-ADS) is an unsupervised deep learning framework to detect accident events and produce compact event-based video summaries from continuous surveillance video streams. The model is trained only with normal traffic videos, and a probabilistic latent representation of typical traffic patterns is learned. At inference time, frames that differ considerably from these learned frames are considered outside the normal range. The overall structure of the proposed CVAE-ADS model is shown in Figure 4.2. It includes the preprocessing stage, the encoder–decoder network pipeline, the anomaly detection pipeline, and the video summarization module.

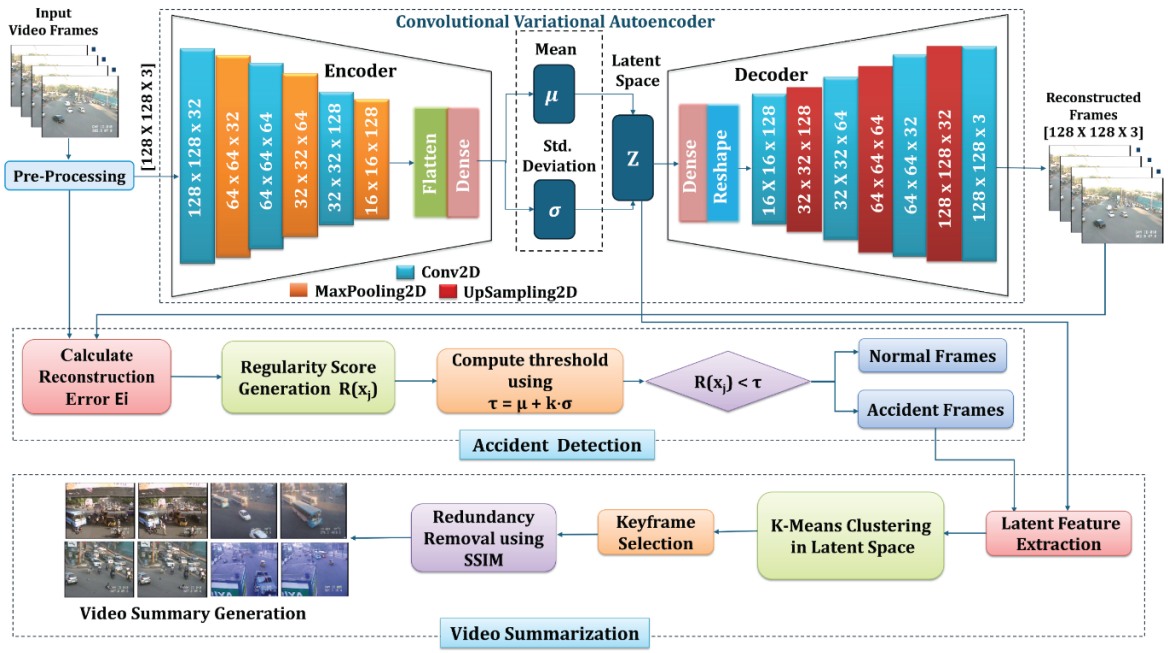


Figure 4.2: Detailed architecture of the proposed CVAE-ADS based Accident detection and video summarization.

4.4.1 Data Preprocessing

The raw frames of surveillance video are first passed through the model after a standardized preprocessing pipeline. All the frames are resized to the same spatial resolution, 128 x 128, and the intensity of each pixel is divided by 255 to have a range of [0,1]. This normalization step helps to ensure consistent input scaling of footage captured from different camera sources, which might be different resolutions, different lighting conditions, different contrast levels, etc. Uniform preprocessing can not only help to maintain a more stable gradient flow during training but also significantly shorten the training time of the model and improve the robustness of the input distribution.

4.4.2 Encoder Architecture

The encoder needs to process each input frame to get the compact and hierarchical spatial representations and to transfer these representations into a probabilistic latent space. The depth of the representations learned and the spatial dimension of the feature maps decrease with the number of layers in the stack, which is composed of a series of convolutional and max-pooling layers. The encoder, in particular, compresses a 128x128x3 frame to a size of 16x16 by feeding it through a series of 32, 64, and 128-sized convolutional layers. These feature maps are flattened and passed through a fully connected (dense) layer to generate the mean (μ) and standard deviation (σ) of the learned latent distribution. This probabilistic

method is not based on a fixed-point vector for every frame; the encoder is able to model the typical traffic distribution.

4.4.3 Latent Space Representation

The latent space is a low-dimensional, probabilistic representation of the input frame that is continuous. CVAE doesn't encode a frame to a fixed point, but rather uses a Gaussian distribution with parameters μ and σ. This distribution is re-parameterized to sample a latent vector z from it and to train end-to-end with gradients [109, 113]:

$$z = \mu + \sigma \odot \epsilon, \ \epsilon \sim N(0,1) \tag{4.1}$$

where ϵ is a standard normally distributed random noise vector. This formulation allows for preserving differentiability through sampling operation, which allows the model to learn a structured and regularized latent space. On normal traffic frames, the latent space will converge to the typical distribution of traffic patterns during training. Therefore, frames with accidents and other anomalies generate latent vectors that are not in this learned distribution and generate poor reconstruction quality and high reconstruction error.

4.4.4 Decoder Architecture

The decoder will send the sampled latent vector z back to the input frame. It starts at a dense layer which projects back the latent vector to a higher dimensional space and then reshapes it to a 16×16×128 three-dimensional feature map. The spatial resolution is built up step by step using a series of up-sampling layers, which are followed by convolutional operations and a progressively increasing number of intermediate feature map sizes of 32×32, 64×64, and 128×128. The filter depths are exactly opposite to the encoder (128, 64, 32), and the last layer is a convolutional layer that gives the reconstructed frame $\hat{x}$ with shape (128, 128, 3). During training, the decoder is told to keep this distance between the input image x and the reconstructed image $\hat{x}$ as small as possible. The distance between original image and the reconstructed image is the main parameter to detect abnormal situations in the inference phase.

4.4.5 Loss Function Formulation

The CVAE-ADS model optimizes a composite loss function that makes use of two complementary objectives:

$$L_{CVAE} = L_{recon} + L_{KL} \tag{4.2}$$

The reconstruction loss L_{recon} is the pixel-wise difference between the original frame x and the reconstructed frame $\hat{x}$ usually the mean squared error (MSE) across all spatial positions and channels. This makes the decoder more likely to reconstruct the input frames that are consistent with the distribution of normal traffic patterns learned.

The KL divergence loss L_{KL} [109, 114] encourages the latent space to be regularized, that is, to push the learnt posterior distribution $q(z \mid x)$ to be close to the standard normal prior $p(z) = N(0, I)$:

$$L_{KL} = -\frac{1}{2}\sum(1 + \log(\sigma^2) - \mu^2 - \sigma^2) \tag{4.3}$$

This regularization helps to avoid the latent space becoming too sparse or broken up, so similar normal frames end up in neighbouring parts of the space, and the model will generalize well to new normal traffic conditions. The combination of these two terms of loss guarantees that the model can learn not only good reconstruction but also a well-structured probabilistic latent representation.

4.4.6 Anomaly Detection and Thresholding

At the time of inference, the trained CVAE-ADS model will reconstruct every incoming surveillance frame. The model has been trained only for normal traffic patterns, so that it has high fidelity in reconstructing normal frames, and frames with anomalous events like accidents do not fall under the learned distribution and are poorly reconstructed. This difference is calculated by the MSE of all the pixels and all the channels [113]:

$$E(x_j) = \frac{1}{HWC}\sum\left(x_j - \hat{x}_j\right)^2 \tag{4.4}$$

where H, W, and C are the height, width, and number of channels, respectively, of the frame. The error is standardized into a regularity score to compare the raw reconstruction error between frames and video sequences [114]:

$$R(x_j) = 1 - \frac{E(x_j) - \min(E)}{\max(E) - \min(E) + \epsilon} \tag{4.5}$$

where ϵ is a small constant added for numerical stability. A value of 1 is considered a normal

frame, and a value of 0 is considered an anomalous frame. Then, an adaptive threshold τ is calculated on the basis of the statistical distribution of the reconstruction errors:

$$\tau = \mu_{err} + k \cdot \sigma_{err} \tag{4.6}$$

where k is a sensitivity parameter, and μ_{err} and σ_{err} are the average and standard deviation of all reconstruction errors.

$$\text{Frame Classification} = \begin{cases} \text{Normal, if } R(x_j) \geq \tau \\ \text{Accident, if } R(x_j) < \tau \end{cases} \tag{4.7}$$

The frames that are detected as anomalous are fed to the video summarization framework described in Section 4.6, whereas the normal frames are not considered for any further processing.

4.5 Proposed Conv-BiLSTM Autoencoder Framework

Both spatial characteristics and temporal dependencies are present in traffic video sequences, and the proposed Conv-BiLSTM Autoencoder framework is capable of effectively modelling them. Unlike traditional methods, which consider each frame separately, this framework makes use of sequential learning to understand the motion and the evolution of the context over time. The system is an integration of convolutional feature extraction and

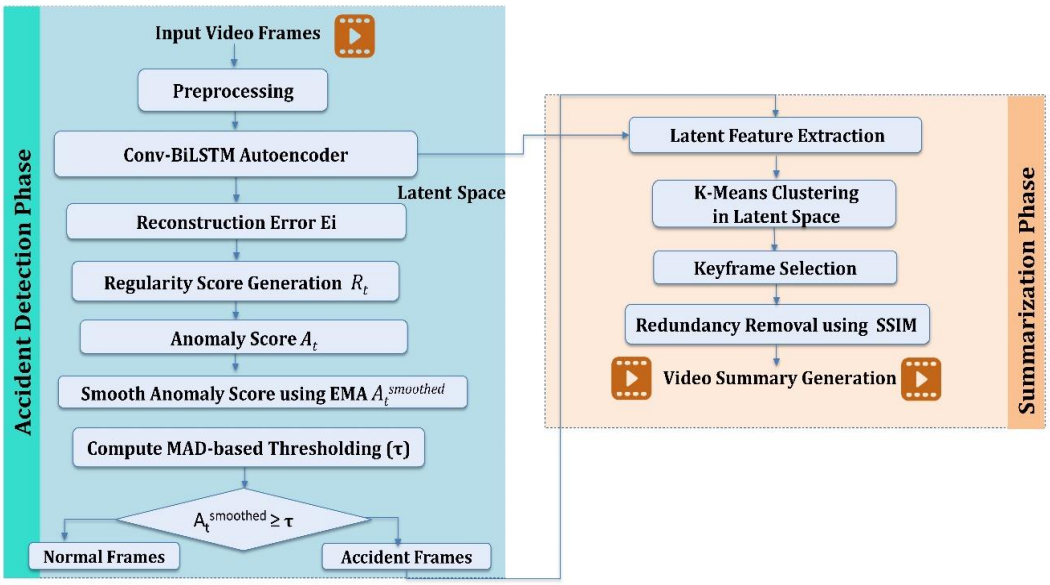


Figure 4.3: Block Diagram of Proposed Conv-BiLSTM Autoencoder Framework for Accident Detection and Video Summarization

bidirectional temporal modelling with a reconstruction-based anomaly detection mechanism and an event-driven video summarization pipeline, as shown in Figure 4.3 and Figure 4.4.

The architecture has two main phases:

1. Accident Detection Phase: uses reconstruction error to detect abnormal events.
2. Summarization Phase: extracts meaningful keyframes to represent detected events.

Each aspect of the framework is detailed in the following subsections.

4.5.1 Input Preparation and Preprocessing

The consistency and quality of the input data have a significant impact on the deep learning model's performance. Hence, a preprocessing pipeline is used to process raw traffic videos to obtain appropriate input sequences for the Conv-BiLSTM autoencoder.

In the first step, the continuous video streams are split up into individual frames at a fixed frame rate. The frames are extracted, and each frame is resized to a uniform size (e.g., 128 × 128 pixels) for computational efficiency and to be compatible with convolutional layers. The pixel values can be rescaled between 0 and 1 to make the flow of the gradients more stable and train faster.

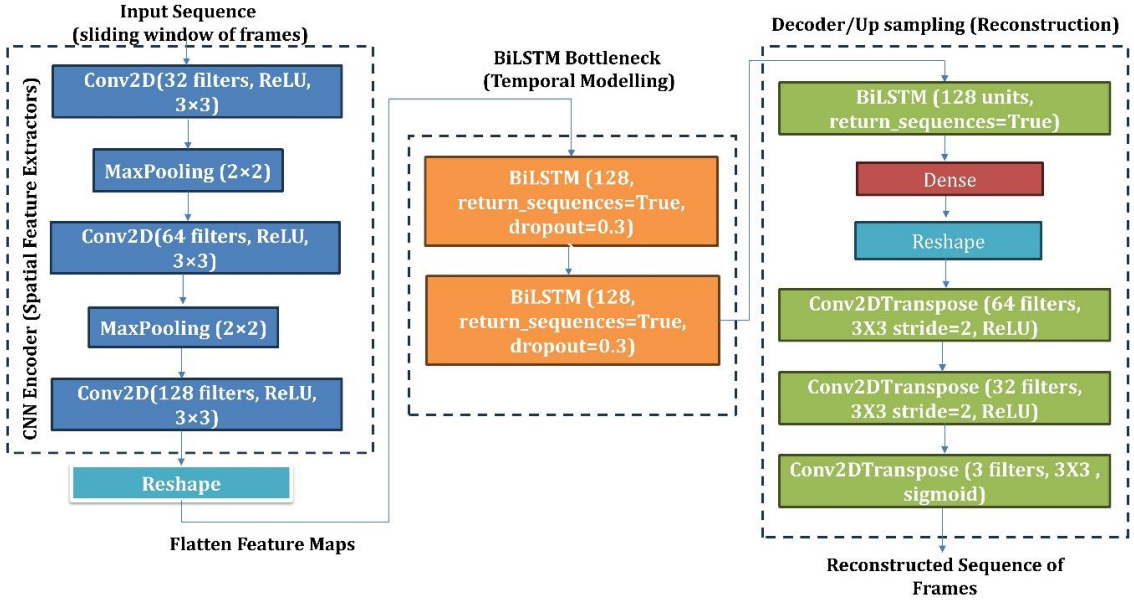


Figure 4.4: Architectural diagram of the proposed Conv-BiLSTM Autoencoder-based model

A sliding window approach is used to group several frames into fixed-length sequences, capturing temporal relationships. Suppose that the length of a sequence is T, then each sample entering the sequence can be expressed as:

$$X = x_1, x_2, x_3, \dots, x_T \tag{4.8}$$

where each x_t representing a pre-processed video frame. This temporal order allows the model to capture the temporal structure of the movements as well as the spatial structure.

Furthermore, optional preprocessing steps can be used, for example, to filter out noise and normalize the contrast to enhance the generalisation and robustness. This model, however, is mostly focused on learning the regular patterns directly from the data without extensive preprocessing that is handcrafted.

4.5.2 Spatial Feature Extraction

Next, the spatial features from each frame are separated through a series of convolutional layers, and then the encoder begins. The convolutional operations are applied time-distributed, which means that every frame of the sequence is processed in the same way.

The encoder is made up of 3 convolutional blocks, each with a deeper filter:

- The first convolutional layer has 32 filters of size 3 x 3 with the ReLU activation function.
- The second convolutional layer has 64 filters, each 3 x 3 in size and activated by the ReLU function.
- The 3rd convolutional layer has a 3 x 3 kernel size, 128 filters, and ReLU activation.

After each convolution layer, there is a max pooling (2x2) which helps to retain important structural information and to reduce the spatial size. The model is able to learn high-level properties such as object shape and scene context as well as low-level properties such as edges and texture.

In mathematical terms, the convolution operation is:

$$F_t = \sigma(W * x_t + b) \tag{4.9}$$

where σ is the ReLU activation, W is the convolving weights, b is a bias, and F_t is a feature map at time t.

The feature maps from the convolutional layers are then flattened and sorted into a sequence of features for temporal modelling. In this transformation, spatial representations are turned to a temporal feature vector sequence:

$$Z = z_1, z_2, z_3, \dots, z_T \tag{4.10}$$

where each z_t contains the spatial information from each frame x_t.

4.5.3 Temporal Modelling using BiLSTM

The framework is based on Bidirectional Long Short-Term Memory (BiLSTM) networks to model temporal dependencies in both forward and backward directions. The difference between standard LSTM and BiLSTM is that BiLSTM takes both the past and future context into account, which is helpful for analysing videos [103] [120].

The flattened feature sequence is fed through bi-directional LSTM (BiLSTM) layers, with 128 hidden units in each layer. These layers acquire complex temporal relationships, such as continuity of movement, abrupt changes, and how events transition.

The BiLSTM calculates the following for a given time step t:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{4.11}$$

Where $\overrightarrow{h_t}$ represents the forward output of LSTM and $\overleftarrow{h_t}$ Represents the backward output of LSTM.

The string of these outputs provides a complete description of temporal context. A dropout regularization is used with a 0.3 dropout rate for preventing overfitting and enhancing generalization. The output of the BiLSTM layers constitutes a compact latent representation (also known as the bottleneck of the autoencoder). This latent space is composed of Spatial structure (CNN encoder) and Temporal evolution (BiLSTM).

4.5.4 Encoder–Decoder Structure

The proposed Conv-BiLSTM autoencoder has an encoder-decoder architecture to learn a compact representation of the input video sequences and then reconstruct them. This is because the decoder tries to reconstruct the original sequence from the "compressed" latent space, and the encoder tries to compress the high-dimensional information of the spatiotemporal data into the reduced-dimensional latent space.

The input sequences are fed to the encoder through the time-distributed convolutional layers, which make sure that the spatial features are extracted in the same way from every frame. These convolutional processes, paired with max-pooling layers, decrease the spatial dimension of the feature maps but keep the most informative features. Then, the feature maps are reshaped back into a sequential format for temporal modelling. Stacked BiLSTM layers are used to model the dynamic dependencies between frames. This Spatio-temporal processing capability of the encoder enables it to encode both the spatial structure and temporal evolution into a single latent representation.

During decoding, the encoder is able to reconstruct the original sequence by utilizing the latent features learned during the encoding process. It starts with a BiLSTM layer to obtain sequences as output, thus maintaining the temporal continuity when reconstructing. These temporal features are then fed into a fully connected layer and reshaped to output the spatial dimensions needed for image reconstruction. Then a sequence of transposed convolution (Conv2DTranspose) layers is applied to gradually upsample the feature maps. These layers progressively sharpen the output image that is being reconstructed, from coarse to detailed, frame-level structures. The last reconstruction layer uses a sigmoid activation function to keep the reconstructed pixel intensities within the normalized range, generating the realistic RGB frames.

In summary, this encoder–decoder architecture sets up an efficient mapping between the input sequences and their reconstruction. The combination of spatial and temporal modelling allows for accurate reconstruction of normal patterns and also for revealing deviations pertinent to anomaly detection.

4.6 Accident Detection using Conv-BiLSTM Framework

The accident detection mechanism is formulated as a reconstruction-based anomaly detection problem, in which the deviation between input sequences and their reconstructions produced by the proposed Conv-BiLSTM autoencoder serves as the basis for detecting abnormal events. The model is trained exclusively on normal traffic sequences, enabling it to learn the underlying spatio-temporal distribution of regular patterns: the convolutional layers capture spatial information such as vehicle structure, road geometry, and motion boundaries, while the Bidirectional LSTM layers model temporal dependencies in both forward and backward directions. This unsupervised formulation is particularly practical for real-world deployment, where annotated accident data is limited and highly imbalanced.

During inference, sequences that deviate from the learned distribution, due to sudden collisions, irregular trajectories, abrupt changes in object motion, or unexpected object appearances, yield higher reconstruction discrepancies. These deviations are converted into measurable statistical signals that can be systematically analysed for anomaly detection. Moreover, sequence-based modelling incorporates temporal context that is unavailable in frame-level analysis, which is especially important in traffic scenarios where anomalies are inherently temporal phenomena. The following subsections (4.6.1–4.6.5) describe how the reconstruction error, regularity score, anomaly score, temporal smoothing, and thresholding are computed.

4.6.1 Reconstruction Error Analysis

The trained model produces a reconstructed sequence that is close to the original frames for each input sequence. The reconstruction process tries to keep this difference between the input and output representations to a minimum, and the latent space learnt is a compressed one that captures the essential nature of normal traffic patterns.

Reconstruction quality is measured by Mean Squared Error (MSE), which is calculated at each time step between the input frame x_t and its reconstruction $\hat{x}_t$. This error is a measure of the model's performance in normal pattern representation; the lower, the more accurately the model can reconstruct the normal patterns, and the higher, the more there are irregular or unseen events.

$$E_t = \| x_t - \hat{x}_t \|^2 \tag{4.12}$$

The main signal for anomaly detection is the sequence of reconstruction errors. Theoretically, in the training process, the autoencoder only minimizes the reconstruction loss for the normal samples. Thus, if an abnormal sequence is encountered during inference, the latent representation will not be able to generalize, resulting in greater reconstruction error. This property allows the model to implicitly estimate the normal data distribution as a density estimation model.

Also, the errors in reconstruction can be understood as likelihood under the learned manifold of normal behaviour. The error values are much higher for frames outside this manifold, for example, if an accident occurs. This renders reconstruction error as a compelling and intuitive measure of abnormality in a video sequence. Reconstruction errors can be summed over spatial dimensions (pixel-wise) and temporal windows to further increase robustness, thus giving a more stable representation of anomaly intensities throughout the sequence.

4.6.2 Regularity Score Computation

The raw reconstruction errors can be different from sequence to sequence and under different video conditions, so errors are normalized to make them bounded and comparable. The error values are normalized over the sequence, and a regularity score is computed to ensure a consistent interpretation:

$$R_t = 1 - \frac{E_t - \min(E)}{\max(E) - \min(E)} \tag{4.13}$$

This formulation is such that those frames which are close to normal patterns are given higher scores, and those frames which are irregular or abnormal are given lower scores. It is important to note that the normalization process to rescale the reconstruction error into a relative measure of normality within a given sequence is an effective process. This is especially convenient when considering video footage taken under different environmental conditions, and makes it less sensitive to absolute error magnitudes.

Furthermore, the regularity score provides interpretability to the detection system. Rather than using raw error data, which can be difficult to interpret, the normalised score gives a good indication of whether each frame is similar to or different from the normal behaviours learned. This transformation also helps in subsequent processing steps like thresholding, temporal smoothing, etc., since the range of all sequences is now in a consistent range of numbers.

4.6.3 Anomaly Score Calculation

The regularity score is converted into an anomaly score by taking the inverse of the regularity score to directly measure the level of abnormality. This transformation makes interpretation easier because the higher the value, the higher the likelihood of the anomaly:

$$A_t = 1 - R_t \tag{4.14}$$

Thus, the frames that have substantial inconsistencies in the frame reconstruction process have higher scores of anomalies and may be considered as possible accident events.

Anomaly score is a direct indicator of deviation from normal behaviour. The framework makes regularity an abnormal situation, which makes the scoring mechanism more in line with the goal of capturing rare and unusual events.

The anomaly score sequence can be treated as a time-series signal from a signal processing point of view, and peaks of the signal are identified as abnormal events. This representation can be used to apply temporal filtering and statistical analysis techniques to enhance detection performance.

In addition to that, the anomaly score allows for the easy integration with other decision-making strategies, like adaptive thresholding, clustering, or event segmentation. It also enables the easy visualization of the abnormal patterns over time, which is helpful for evaluation and interpretability.

4.6.4 EMA Smoothing

There can be short-term changes in the anomaly score sequence because of illumination changes, motion variations, or noise. To make the results more temporally stable and reduce false positives, results are computed using an Exponential Moving Average (EMA):

$$S_t = \alpha A_t + (1 - \alpha) S_{t-1} \tag{4.15}$$

where α is the smoothing factor which determines the weight of the latest observations.

EMA smoothing is a low-pass filter that reduces high-frequency noise and gives good, meaningful temporal trends. It reduces the sensitivity to transient disturbances while keeping responsiveness to the real anomalies in accordance with the more recent observations. The

smoothing parameter α is an important parameter to consider when balancing sensitivity and stability. The higher the value of α the more recent changes will be emphasized, and the quicker sudden anomalies can be detected, but the lower the value of α the smoother the signals get, the more anomalies will be delayed in detection.

This is especially crucial in a traffic environment where environmental changes like lighting changes, shadows, or slight motion of objects can create noise in the anomaly signal.

4.6.5 MAD-Based Thresholding

A strong threshold is determined by Median Absolute Deviation (MAD), which is not as significantly affected by outliers as mean-based statistics. Let m be the median of the smoothed anomaly scores. Threshold is defined as:

$$\tau = \mathrm{m} + \mathrm{k} \cdot \mathrm{MAD} \tag{4.16}$$

where k is a scaling constant that determines the detection sensitivity. Sequences or frames that meet the criterion:

$$S_t \geq \tau \tag{4.17}$$

High values of S_t are considered accident events. Because MAD is robust to skewed distributions and extreme values, it will not be greatly affected by a few outlying frames on the threshold. This is essential for actual video data, where there is noise and variability.

Moreover, thresholding is adaptive in the MAD-based method, which means that manually defined thresholds are not required to adapt to different video conditions. This improves the generalization and renders the method applicable to various surveillance scenes.

The parameter k can be adjusted to balance false positives and false negatives. A more stringent criterion for detection will be obtained with a larger value of k, though the sensitivity to detect subtle anomalies would be reduced, and a less stringent criterion will be obtained with a smaller value of k, though there would be more false detections.

4.7 Video Summarization Framework

This module aims to obtain a short and informative summarization of the detected accident events by choosing a subsequence of representative frames from the anomaly sequences. In

large-scale traffic surveillance systems, continuous video streams produce huge data volumes, and it is impractical to manually inspect all of the videos. Hence, it is crucial to have an automatic summarization system to remove the need to store redundant data and to obtain important information.

The proposed approach is different from the conventional frame sampling and uniform subsampling methods that usually neglect semantic relevance and are performed in latent feature space. This allows semantically grouping of frames using representations learned from the data, instead of only using pixel-level similarity. The deep feature embeddings guarantee that visually and contextually similar frames are clustered together, allowing to improve the quality of the generated summary.

The summarization framework proposed can be used with both CVAE-ADS and Conv-BiLSTM autoencoder models. The proposed CVAE-ADS encoder learns compact spatial representations that include appearance features, including event evolution, object structure, and scene layout, while the proposed Conv-BiLSTM encoder learns enriched Spatio-temporal features that include both appearance and temporal features such as motion continuity and event evolution. The dual compatibility makes the framework work for various feature representations without requiring any structural changes.

The summarization process has five stages, namely, latent feature extraction, clustering, keyframe selection, redundancy removal, and the generation of a final summary. Each step will be used to progressively remove redundant information from the events while still keeping necessary information about the events so that the result will be concise and informative.

4.7.1 Latent Feature Extraction

After detecting the anomalous sequences (as mentioned in Section 4.5), the detected frames are fed to the encoder of the trained model (CVAE-ADS or Conv-BiLSTM) to get the latent representations of the frames. The latent vectors give a compact and discriminative encoding of the visual content. Let the set of detected anomalous frames be $\{x_1, x_2, \dots, x_T\}$. A latent vector is created for each frame:

$$z_t = f_{enc}(x_t) \tag{4.18}$$

where $f_{enc}(\cdot)$ is the encoder function and $z_t \in R^d$ Is the latent embedding.

The resulting latent sequence is $Z = z_1, z_2, \ldots, z_T$. High-level structure and context information are captured and can be used for clustering and similarity analysis. Latent space can be considered a compressed manifold for representation learning, in which semantically similar frames are located near each other. This is an important property to have for effective summarization because it groups by content and not just by pixel differences.

Furthermore, latent representations are more resistant to noise, lighting variations, and small spatial variations than raw pixel values. This robustness makes it possible for frames showing the same event with some variation of the circumstances to become clustered together.

In the Conv-BiLSTM encoder, the latent vectors are implicitly temporal, that is, every latent vector represents the current frame but also takes into account information from preceding and following frames. In this way, the quality of summarization is improved with the property of temporal coherence.

4.7.2 K-Means Clustering

The similar frames are clustered using K-Means so as to reduce the redundancy of latent vectors. In order to maximize similarity inside a cluster and minimize similarity between clusters, the latent space is divided into K clusters [70].

The latent vectors are clustered by K-Means clustering in order to group similar frames and reduce the redundancy. The goal is to divide the latent space into K clusters, maximizing the similarity within a cluster while minimizing the similarity between clusters [70].

$$\min_{\{C_k\}} \sum_{k=1}^{K} \sum_{z_i \in C_k} \|z_i - \mu_k\|^2 \tag{4.19}$$

Where the k^{th} cluster is denoted by C_k, and its centroid is denoted by μ_k.

Each cluster represents a set of frames that share similar visual and contextual characteristics. Clustering in the latent space has multiple benefits compared to pixel space clustering. Firstly, it decreases the dimension, thus reducing the computational complexity. Second, it guarantees that the decisions about clustering are made based on the semantically meaningful features learned by the model.

The value of K is chosen adaptively, according to the number of detected anomalous frames, so that the summary is compact, while keeping information. A low value of K will make the summary more compact but may lose valuable variations, while a high value of K Will keep more detail but introduce more redundancy. An anomaly-proportional approach is used to determine the number of clusters, namely K is scaled according to the number of detected anomalous frames instead of the sequence length. This way, clustering will adapt to the complexity of the event and not the length of the video. Experimental results show that this approach is stable to generate summaries with minimal redundant information while keeping the key event transitions and is thus suitable for accident-based video summarization where the event dynamics are very different from one video sequence to another.

4.7.3 Keyframe Selection

A summary is created from a representative frame from each cluster. The best frame for the keyframe is the closest latent representation to the cluster centroid.

$$x_k^* = \arg\min_{x_i \in C_k} \| z_i - \mu_k \| \tag{4.20}$$

This selection strategy assures that keyframes well represent the visual content of each cluster. The centroid-based selection mechanism is based on the hypothesis that the centroid is the average of all frames in the cluster. Thus, the frame that is the closest to the centroid represents the best sample of a group.

The method is to choose one frame from each cluster to provide various aspects of the accident event: before the event, at the moment of collision, and after the collision. This variety provides the freedom that the created overview of the event does not tend to a particular time, but gives an overview of the event as a whole. Moreover, this method does not pick outlier frames, which might be chosen by considering only the reconstruction error or by random sampling.

4.7.4 Redundancy Removal using SSIM

Clustering will decrease the amount of redundancy, but there might be high visual similarity between some of the keyframes that are selected. In order to further reduce the summarization, the Structural Similarity Index (SSIM) [125] is applied as a redundancy elimination step.

The SSIM of two candidate frames x_i and x_j is computed as:

$$\mathrm{SSIM}(x_i, x_j) = \frac{(2\mu_i\mu_j + C_1)(2\sigma_{ij} + C_2)}{(\mu_i^2 + \mu_j^2 + C_1)(\sigma_i^2 + \sigma_j^2 + C_2)} \tag{4.21}$$

Where μ, σ, and σ_{ij} are the mean, variance, and covariance, respectively.

If $\mathrm{SSIM}(x_i, x_j) > \theta$, the frames are regarded as redundant frames, and the other frame is discarded. The threshold $\theta \in [0,1]$ determines how much the redundancy is removed; the greater the value, the smaller the number of frames kept when they are very different from each other. This way, the final summary will not have repeated frames and will contain visually different and informative content. The higher the threshold, the more redundant frames will be removed, the more compact the summary will be; the lower the threshold the more frames will be retained.

Further, SSIM-based filtering can supplement the clustering procedure, as some similar frames might end up in different clusters because of the minor changes in their latent space representation.

4.7.5 Summary Video Generation

The last set of keyframes is sequenced to maintain the order of events in the original timeline of the keyframes. This order preserves the logical sequence of the accident, beginning with the pre-event buildup, when the accident occurred, and the post-event. The ordered frames are then assembled into a summary video at a predetermined frame rate (e.g., 5 fps), which produces a concise yet informative summary of the accidents detected.

Formally, if the key frames selected are $\{x_{k_1}^*, x_{k_2}^*, \ldots, x_{k_n}^*\}$, then the summary sequence is:

$$S = \mathrm{sort}_t(x_{k_1}^*, x_{k_2}^*, \ldots, x_{k_n}^*) \tag{4.22}$$

where sort_t is a function that sorts frames by temporal order.

The proposed summarization framework uses latent feature representation to produce a meaningful and structured representation of accident events. The approach based on clustering-based grouping, keyframe selection based on the centroid, and redundancy removal based on SSIM ensures the diversity and compactness in the generated summary.

Moreover, it is suitable for both CVAE-ADS and Conv-BiLSTM models, providing a single and flexible tool for event-based video summarization. Combining deep representation learning with classical clustering and similarity measures is a good balance between efficiency and semantic richness. In summary, the overall design is a data integrity and event-recordation system that will not only save data redundancy but also preserve the critical event data, which is very suitable to be used in the intelligent surveillance system and real-time monitoring application.

4.8 Algorithmic Workflow of the Proposed System

The entire algorithmic process of the proposed accident detection and event-based video summarization framework is presented in this section. Two distinct models, CVAE-ADS and Conv-BiLSTM autoencoder frameworks, are developed to identify anomalous events and provide a concise summary of accident sequences. To enable the workflow to operate in an unsupervised manner by learning the normal traffic patterns and detecting deviations during inference.

Algorithm 4.1: Accident Detection and Video Summarization Framework

Input: Surveillance video V

Output: Detected accident frames/sequences and summarized video S

1. **Frame Extraction**
   Extract frames from the input video:

$$F = f_1, f_2, f_3, \dots, f_n$$

2. **Preprocessing**
   Normalize the intensity of the pixels in each frame to a fixed resolution (e.g., 128 x 128) to represent the same input.

3. **Input Representation Generation**

   - For **CVAE-ADS**: Process each frame separately.
   - For **Conv-BiLSTM**: Build frame sequences by sliding window of size T:

$$X = f_t, f_{t+1}, \dots, f_{t+T-1}$$

4. **Model Training**

   Use only normal traffic data to train the selected model, so that it can learn the normal spatial as well as Spatio-temporal patterns.

5. **Reconstruction (Inference Phase)**
   For each input frame/sequence:
   - Use the trained model to pass input and get output
   - Get reconstructed output $\hat{f}_t$ or $\hat{X}$

6. **Reconstruction Error Computation**
   Measure reconstruction error using MSE:

$$E_t = \| f_t - \hat{f}_t \|^2$$

7. **Score Transformation**
   Convert reconstruction error into a normalized measure:
   - Compute regularity score R_t
   - Calculate anomaly score A_t

8. **Anomaly Detection via Thresholding**
   Using model-specific criteria, classify frames/sequences as normal or abnormal:
   - **CVAE-ADS:**
     Compute statistical threshold:

$$T_{CVAE} = \mu + k \cdot \sigma$$

     A frame is considered anomalous if:

$$R_t < T_{CVAE}$$

- **Conv-BiLSTM:**
  Use temporal smoothing (EMA) and thresholding using MAD:

$$T_{BiLSTM} = m + k \times MAD$$

  A sequence is considered anomalous if:

$$S_t \geq T_{BiLSTM}$$

9. **Extraction of Anomalous Segments**

   Detect and locate frames/sequences that are different from the rest, according to the following detection criteria:
   - For **CVAE-ADS**, select frames satisfying:

$$R_t < T_{CVAE}$$

   - For **Conv-BiLSTM**, select sequences satisfying:

$$S_t \geq T_{BiLSTM}$$

These selected frames/ sequences are then passed to the next summarization stage.

10. **Latent Feature Extraction**
    Feed the detected anomalous frames into the encoder to get a latent representation of the frames:

$$z_t = f_{enc}(f_t)$$

11. **Clustering of Events**
    Use K-Means clustering to cluster similar events based on latent vectors.

12. **Keyframe Selection**
    Pick representative frames from each cluster, according to the distances from cluster centroids.

13. **Redundancy Removal**
    Filter out frames that are similar in appearance with SSIM:

$$SSIM(x_i, x_j) > \theta$$

14. **Chronological Ordering**
    Organize selected keyframes by their temporal order.

15. **Summary Generation**
    Create an ordered list of keyframes and build the final, summarized video:

$$S = k_1, k_2, \ldots, k_m$$

We propose two different frameworks for accident detection and video summarization based on events. The first framework uses a CVAE-ADS to learn spatial irregularities at the frame level and reconstruction-based abnormalities, while the second framework models temporal dynamics across frame sequences using a Conv-BiLSTM autoencoder. The two methods are both unsupervised, learning typical traffic behaviours and detecting outliers at test time.

Latent feature representations allow efficient clustering or keyframe selection due to the compact and semantically meaningful description of detected events. The framework clusters together similar abnormal patterns and removes the redundant frames to produce concise summaries without loss of information. The system's overall performance, characterized by the accuracy in detecting accident events and the interpretability provided by the latent space-based summarization, demonstrates the efficacy of the unsupervised anomaly detection approach.

Chapter 5: Experimental Results and Performance Evaluation

5.1 Hardware and Software Environment

The proposed event-based video summarization and accident detection framework was experimented with using a GPU-enabled computing platform, which provided efficient handling of large video data and deep learning workloads. The system configuration facilitated stable training, quick convergence, and reliable evaluation of training performance.

The system used an Intel Core i7 processor (10^{th} Generation) with a base clock speed of 2.6 GHz, ensuring efficient processing of preprocessing steps like frame extraction, normalization, and data handling. To speed up the model training, an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM was used. The parallel processing of the convolutional and recurrent operations in CVAE-ADS and Conv-BiLSTM helped to significantly improve the computational efficiency using the GPU.

The system was equipped with 16 GB DDR4 RAM, which was sufficient for processing high-dimensional video frames, performing batch processing, and intermediate feature representations while training. Moreover, a 512 GB SSD storage was employed to store the datasets, trained model checkpoints, logs, experimental outputs, etc., which will lead to increased data access speed and lowered I/O latency.

The hardware infrastructure was capable of continuous training for several epochs without impacting performance and enabled efficient experimentation with various model configurations and hyperparameters. This setup provided stable model convergence, less training time, and efficient processing of large-scale surveillance data like IITH and UCF-Crime. A cutting-edge deep learning software stack was adopted for the implementation of the proposed framework, which enables flexible model development, efficient computations, and end-to-end video processing.

The main development was done in Python (version 3.9). The design and training of the deep learning models were done in TensorFlow (version 2.10) using the Keras API, which offers an efficient implementation of kernel architectures such as convolutional, variational, and recurrent that were utilized in the CVAE-ADS and Conv-BiLSTM Autoencoder models.

Various libraries were used at different components of the pipeline. Numerical computations and handling of data were performed using NumPy and Pandas. Video preprocessing was done using OpenCV (version 4.6), which includes extracting frames, resizing, and normalizing them. Clustering, evaluation metrics, and dimensionality reduction were performed using Scikit-learn (version 1.1). The image quality measures (SSIM and PSNR) were calculated using the scikit-image library, and the training and performance results were plotted using Matplotlib and Seaborn. To train the model efficiently and make sure of the model training speed, the experiments were performed in a GPU-enabled environment. To ensure the reproducibility and consistency of experiments, standard practices like model checkpointing and logging were used.

5.2 Dataset Description

The proposed framework was evaluated on two publicly available benchmark datasets that are widely used for anomaly detection and accident recognition in surveillance video. Both data sets show a variety of real-world traffic situations and offer a different level of complexity in terms of the dynamics of the scenes, lighting conditions, and the frequency of events. The chosen datasets provide a diverse and balanced assessment of the models proposed in this study, the CVAE-ADS and Conv-BiLSTM Autoencoder, in various surveillance environments.

5.2.1 IITH Dataset

The IIT Hyderabad (IITH) Dataset [119] is a collection of traffic surveillance videos that were recorded using CCTV cameras from various locations in the city of Hyderabad, India.

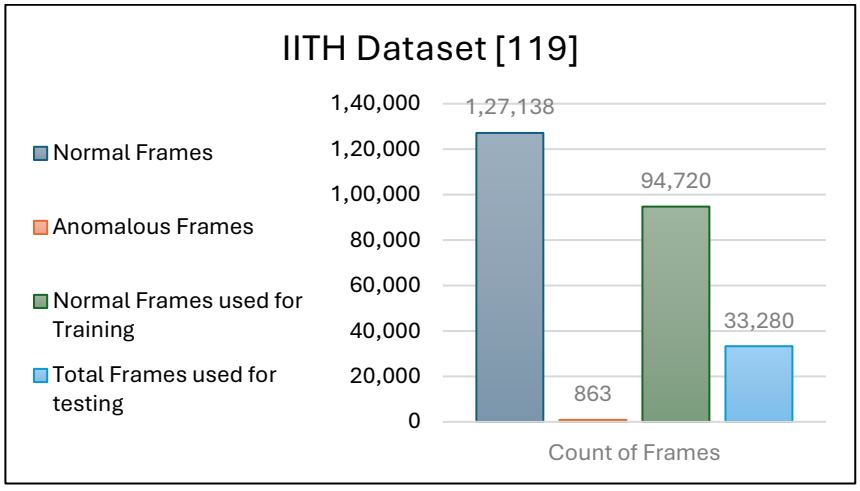


Figure 5.1: Details of IITH Road Accident Dataset

The videos are captured at 30 frames per second and are representative of actual road traffic situations. The typical video (cyber-video) starts with regular traffic flow, and then a traffic accident event occurs. This natural property of the data makes it well suited for unsupervised learning because the model can be trained using only the normal traffic patterns and not involving labelled anomaly data.

The statistics of the dataset, such as the distribution of normal and abnormal frames used in the training and testing, are shown in Figure 5.1. A total of 94,720 normal frames were used for training, and 33,280 frames were used for testing, comprising 32,417 normal frames and 863 accident frames. Also, the representative sample frames of normal traffic and accident states are displayed in Figure 5.2. The fact is that in this study, only the beginning parts of the videos were used for the models' training, where traffic behaviour was normal, thus allowing the models to learn their underlying distribution. Those segments, which included accident events, were for testing and evaluation. This separation guarantees that anomalies are discovered by deviations from learned typical patterns.

Figure 5.2: Video frame samples of the IITH Road accident dataset [119]

5.2.2 UCF-Crime Dataset

The UCF-Crime dataset [128] is the most popular huge real-world dataset for video anomaly detection. It contains about 1,900 long, uncropped videos in 13 categories of abnormal activity, such as abuse, arson, arrest, road accidents, assault, burglary, fighting, explosion, shooting, robbery, shoplifting, vandalism, and stealing, as well as normal activity footage. This dataset contains almost 128 hours of video recordings from a variety of real-world

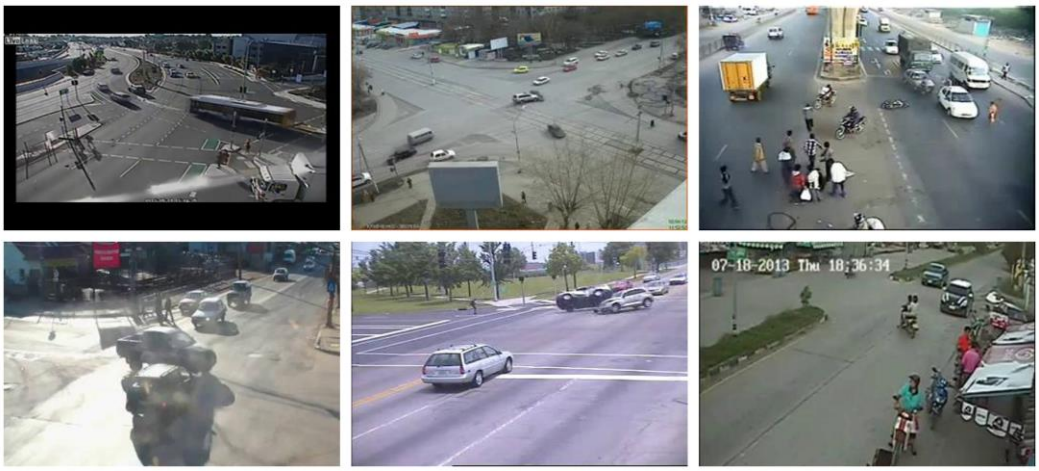


Figure 5.3: Video frame samples of the UCF-Crime Road accident dataset

settings. Figure 5.3 shows representative sample frames from this (Accident subset) dataset showing normal traffic conditions and accident scenarios.

To keep with the objective of traffic accident detection, only the road accident subset (151 videos) was used in this study. From the chosen videos, 127 were used for training, while 23 were used for testing [128] to ensure the proposed models are fairly and consistently evaluated. These videos provide more challenging real-world scenarios, including occlusions, illumination variations, dynamic backgrounds, and varying camera viewpoints, which add complexity and realism to the anomaly detection challenge.

The UCF-Crime dataset does not have predefined frame-level annotations or explicit train-test splits as the IITH dataset. Hence, a preprocessing technique was used where videos were split into individual frames. The fact that the UCF-Crime dataset is large-scale, diverse in scenarios, and complex in visual conditions makes it a good benchmark to assess the generalization capability and real-world applicability of the proposed CVAE-ADS and Conv-BiLSTM Autoencoder framework.

5.3 Data Preprocessing and Training Strategy

The proposed framework is based on an unsupervised learning paradigm, training the models only with normal traffic patterns and using the deviation from the learned patterns to detect anomalies. To guarantee that both datasets were represented consistently as input and that they could be evaluated reliably, a uniform preprocessing pipeline was used to process both.

- **Data Preprocessing**

First, all the surveillance videos from both the IITH and UCF-Crime datasets were decomposed into individual frames at their native frame rate of 30 fps, with every frame retained to preserve complete temporal information. This frame-level representation facilitates the handling of data and matches reconstruction-based learning well. All the frames were resized to a fixed resolution of 128 × 128 × 3 and normalized into [0, 1] to stabilize the training and make it compatible with the sigmoid-based output layer in reconstruction.

The temporal structure of videos was used for the dataset from IITH. Each video usually starts with regular traffic, then it has some accident occurring, so only the first part of the video with normal frames was used for training. All the other frames with accident events were left for testing. Besides, frames were manually labelled as normal or accident for evaluation purposes so as to conduct quantitative performance assessment.

The UCF-Crime dataset also provides temporal information that specifies the temporal interval of abnormal events occurring in video clips. Therefore, manual labelling is not required for this dataset. With the use of these annotations, one may immediately and reliably assess the model's detection abilities on unrestricted real-world video data.

The overall dataset was partitioned into approximately 70% training and 30% testing data. Importantly, the split was performed at the video level rather than the frame level, ensuring that no frames belonging to the same video appear in both the training and testing sets, thereby preventing any data leakage. Consistent with the unsupervised learning paradigm, the training set contained only normal (accident-free) frames, while the testing set contained a mixture of normal and accident frames. For the IITH dataset, this was achieved by using the initial accident-free portions of the videos for training, while the segments containing accident events, along with additional normal segments, were reserved exclusively for testing. Furthermore, 10% of the training data was held out as a validation set, which served two purposes: (i) monitoring the validation loss for early stopping during training, and (ii) estimating the distribution of reconstruction errors on normal data for threshold selection, as described in the anomaly detection stage.

- **Training Strategy**

The CVAE-ADS model is a hybrid of the convolutional and probabilistic latent models used to learn the spatial representations of normal traffic frames. In training, the individual frames

are fed into a convolutional encoder, which is used to generate hierarchical features and map the features to a latent distribution with mean and variance parameters. A latent vector sampled from the input frame (reparameterization trick) is then fed to a convolutional decoder to produce the input frame.

The model is trained with an Adam optimizer with an initial learning rate of 0.0001 to ensure the convergence of the reconstruction and the latent regularization objectives. A batch size of 64 is used in order to achieve a compromise between computational efficiency and learning stability. Training is done until 100 epochs, and validation loss is used to determine early stopping of training. To help regularize the latent space and ensure a smooth representation of the typical traffic pattern, the reconstruction loss (MSE and Kullback-Leibler divergence) is combined in the loss function.

The Conv-BiLSTM Autoencoder on the other hand, aims to learn temporal and spatial properties of traffic scenarios. This is a model that processes sequences of consecutive frames, with a sliding window approach. Specifically, each input sample consists of 16 consecutive frames, and the sliding window is advanced with a stride of one frame, resulting in an overlap of 15 frames between successive sequences. This dense overlapping strategy maximizes the number of training samples generated from the available videos and enables the model to observe smooth temporal transitions of normal traffic flow, which improves the stability of temporal feature learning. In order to learn temporal dependencies in both forward and backward directions, the convolutional layers first extract the spatial features, which are then supplied to the stacked bidirectional LSTM layers. This aids in the model's learning of the typical traffic flow's movement and temporal context.

Since sequence modelling is more complex, a lower learning rate of 0.0001 is used for training the Conv-BiLSTM Autoencoder with the Adam optimizer to ensure a stable flow of gradients through the recurrent layers. The batch size is smaller (16) because of the increased memory requirements. The model is trained for a maximum of 150 epochs, using validation performance for early stopping. To avoid overfitting and enhance its generalization ability, dropout is used in the BiLSTM layers. The training goal is only the reconstruction loss (MSE) over the whole input sequence. Table 5.1 summarizes the important training parameters and configuration for both proposed models for to ensure reproducibility and comparative analysis.

Table 5.1: Training Configuration of Proposed Models

Parameter	CVAE-ADS Model	Conv-BiLSTM Autoencoder
Input Type	Individual frames (128×128×3)	Frame sequences (sliding window)
Training Paradigm	Unsupervised (normal frames only)	Unsupervised (normal sequences only)
Sequence Window	-	16 consecutive frames
Train / Test Split	70% / 30%	70% / 30%
Training Epochs	100	150
Optimizer	Adam	Adam
Learning Rate	0.0001	0.0001
Batch Size	64	16
Dropout	Not used	0.3 (BiLSTM layers)
Normalization	Input normalization [0,1]	Input normalization [0,1]
Activation Functions	ReLU (hidden), Sigmoid (output)	ReLU (CNN), Sigmoid (output)
Loss Function	MSE + KL Divergence	MSE (sequence reconstruction)
Early Stopping	Patience = 10 (val loss)	Patience = 15 (val loss)
Data Augmentation	Horizontal Flip, brightness adjustment	Consistent augmentation per sequence
Random Seed	42	42

- **Anomaly Detection and Thresholding**

Anomaly detection at inference time follows the thresholding formulation defined in Section 4.4.6. For the CVAE-ADS model, the adaptive threshold in Eq. (4.6) was applied with the sensitivity parameter k = 2, where the mean and standard deviation of reconstruction errors were computed over normal validation frames. This value of k was selected empirically through a sensitivity analysis in which k was varied from 1.5 to 3.0 in steps of 0.5, as reported

in Section 5.5.1, where k = 2 achieved the highest F1-score and the most balanced precision–recall trade-off on both datasets.

For the Conv-BiLSTM Autoencoder, anomaly detection follows the pipeline described in Sections 4.6.1–4.6.5. Sequence-level anomaly scores were temporally smoothed using the Exponential Moving Average defined in Eq. (4.15) with a smoothing factor of $\alpha = 0.3$, which suppresses transient single-frame fluctuations caused by noise, illumination changes, or minor motion variations while remaining responsive to sustained deviations. The detection threshold was then computed using the MAD-based rule in Eq. (4.16) with the scaling constant k = 2, where the median and MAD were calculated over the smoothed anomaly scores of normal validation sequences. The MAD formulation was preferred over the mean–standard deviation rule used for the frame-level model because it is robust to outliers and heavy-tailed error distributions, which commonly arise in sequence reconstruction. The value of k was selected empirically through the same sensitivity analysis procedure in which k was varied from 1.5 to 3.0 in steps of 0.5, where k = 2 provided the best balance between precision and recall.

- **Integration with Summarization**

After anomaly detection, the framework uses the latent representations for video summarization. The features from the latent space are clustered to find representative patterns. Keyframes are then picked from each cluster, and redundancy is eliminated based on similarity measures, meaning the resulting summary will include the most important events, but not repeat them.

5.4 Quantitative Performance Evaluation

The quantitative result is evaluated with the help of two publicly available datasets, namely, the IITH Accident and the UCF-Crime dataset. To evaluate the system, two major factors are considered: (i) accident detection performance and (ii) video summarization effectiveness. To show the effectiveness of the proposed models, CVAE-ADS and Conv-BiLSTM under various real-world conditions, they are compared with the baseline models.

5.4.1 Accident Detection Results

The effectiveness of the proposed accident detection framework is tested on two benchmark data sets: IITH Accident Dataset and UCF-Crime dataset, with the help of standard classification metrics. The results are summarised in Tables 5.2 and 5.3, and visual (side-by-side) comparisons are presented in Figures 5.4 and 5.5.

Performance on the IITH Dataset

The quantitative results obtained in the experiments with the IITH database are evidence of the superiority of the proposed models over the baseline approaches. The baseline CAE and VAE models perform moderately with 86.78% and 83.42% accuracy, respectively, as shown in Table 5.2 and Figure 5.4. They have relatively low recall values, which suggest that they are not able to identify all types of accidents, especially in complex traffic environments.

A significant improvement in detecting capability is observed in the proposed CVAE-ADS model with an accuracy of 92.5%, a precision of 0.885, and an F1 score of 0.851. The resulting AUC value is also enhanced to 0.912, which indicates better separability between the normal and anomalous frames. This improvement is due to the probabilistic latent representation that can better capture normal traffic variations and detect deviations effectively.

The Conv-BiLSTM model further improves its performance with the highest accuracy of 94.0% and a recall of 0.91. The AUC value of 0.946 is indicative of strong discriminative capability, while EER is 14.97%, which is a better balance between false and negative rates.

Table 5.2: 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

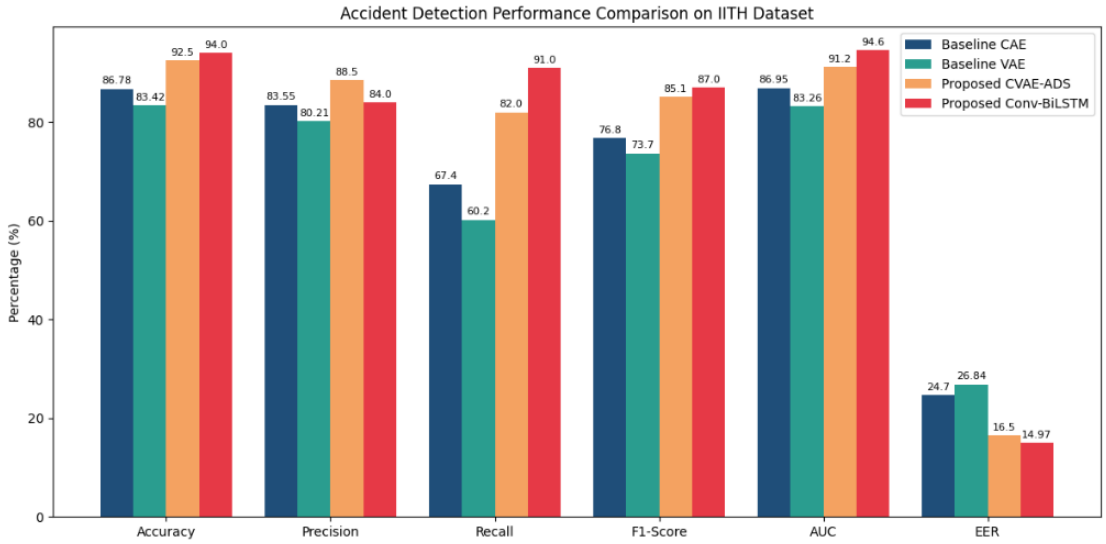


Figure 5.4: Performance Comparison of Accident Detection on the IITH Dataset.

The reason for this is that this model can take temporal dependency into consideration, enabling it to capture the pattern of motion and the context from the consecutive frames.

Performance on the UCF-Crime Dataset

The same performance pattern is seen on the more difficult UCF-Crime dataset, as seen in Table 5.3 and Figure 5.5. The baseline models show less detection capability with 81.63% (CAE) and 78.56% (VAE) and have also lower recall values. This shows that the method is not very successful in locating the scene of the accident with respect to complex real-world scenarios with occlusion, illumination changes, and variance in viewpoints. The proposed CVAE-ADS model significantly outperforms the other models with an accuracy of 88.6% and an AUC of 0.893. The Recall is increasing (0.791), indicating better identification of accident frames than the baseline models.

Table 5.3: 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	0.918	16.20

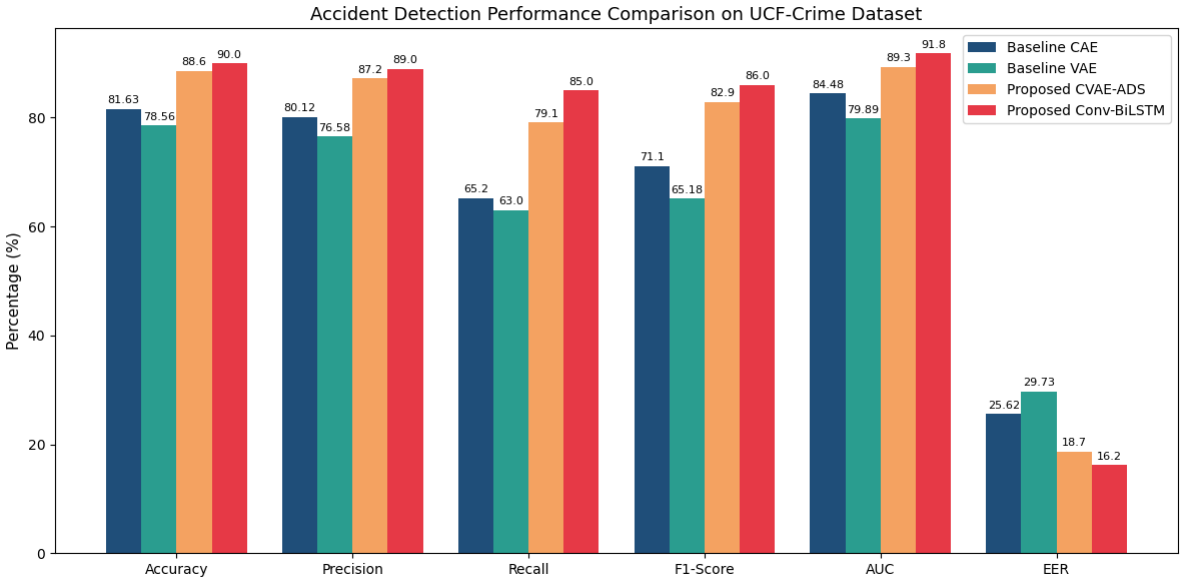


Figure 5.5: Performance Comparison of Accident Detection on the UCF-Crime Dataset

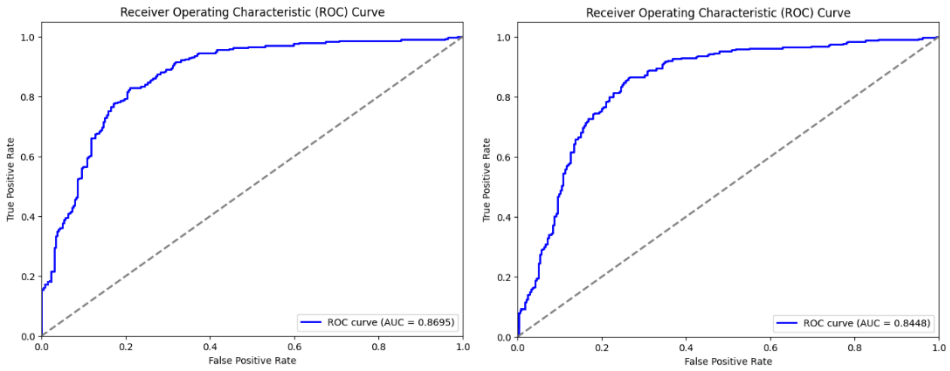
The Conv-BiLSTM model once again provides the best overall performance with an accuracy of 90.0%, recall of 0.85, and AUC of 0.918 with reduced EER of 16.20%. The results show that both the robustness and detection reliability in complex surveillance environments are improved by adding temporal information.

- **ROC Curve Analysis**

Figure 5.6 and Figure 5.7 present the Receiver Operating Characteristic Curves for the baseline models of CAE and VAE, respectively, on the IITH and UCF-Crime datasets. The discriminative ability of these baseline models was moderate, with AUC values ranging from approximately 0.80 to 0.87, which means that these models were not very effective in distinguishing normal and anomalous frames, especially under complex traffic conditions.

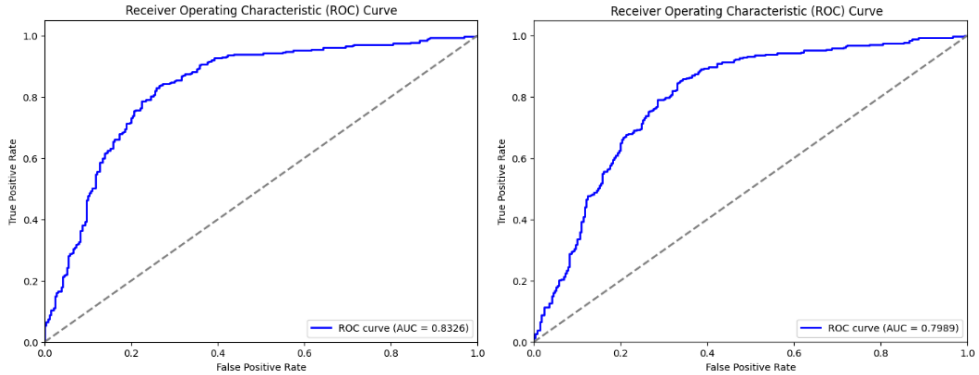
The ROC curves of the proposed models, shown in Figure 5.8 and Figure 5.9, however, show a significant enhancement in detection capability. The CVAE-ADS model obtains 0.912 on IITH dataset and 0.893 on the UCF-Crime dataset, indicating better anomaly detection ability. Results from the Conv-BiLSTM model are further improved and have the highest AUC of 0.9460 and 0.918, respectively.

The proposed models achieved higher TPRs with lower FPRs at various thresholds than did the steeper ROC curves and larger area under the curve. This validates the proposed



(a) **(b)**

Figure 5.6: ROC curve analysis of baseline CAE Model: (a) IITH and (b) UCF-Crime Dataset



(a) **(b)**

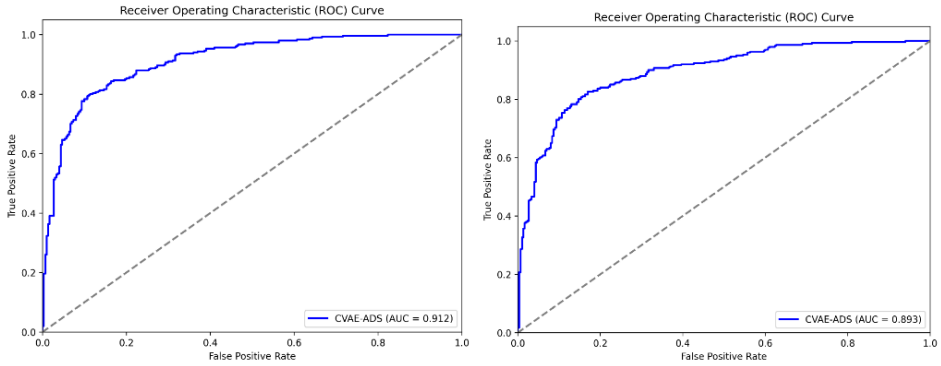
Figure 5.7: ROC curve analysis of baseline VAE Model: (a) the IITH and (b) the UCF-Crime Dataset

approaches and their effectiveness in the accurate detection of accident events with fewer false alarms.

In all the tested models, the Conv-BiLSTM model consistently outperforms the others, as it is able to learn spatial features and temporal dependency by modelling in both directions. In brief, the ROC analysis shows that reconstruction-based unsupervised learning with strong feature representation and temporal modelling can be reliable and practical in real-world traffic surveillance systems for accident detection.

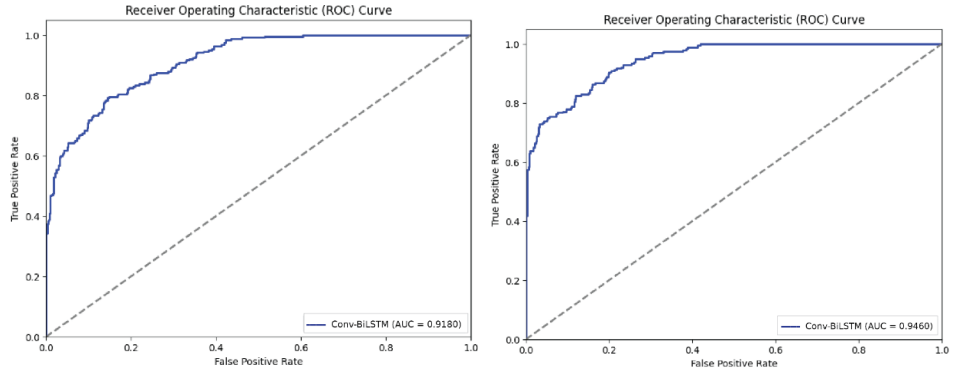
- **Analysis of Confusion Matrix**

The confusion matrices of the proposed models on IITH and UCF-Crime are shown in Figure 5.10 and Figure 5.11, respectively, to give a fine-grained evaluation of the models' classification performance. These matrices provide a detailed breakdown of TPs (true positives), TNs (true negatives), FPs (false positives), and FNs (false negatives), which goes beyond just summarizing metrics. The confusion matrices are generated using the entire test



(a) **(b)**

Figure 5.8: ROC curve analysis of CVAE-ADS Model: (a) the IITH and (b) the UCF-Crime Dataset



(a) **(b)**

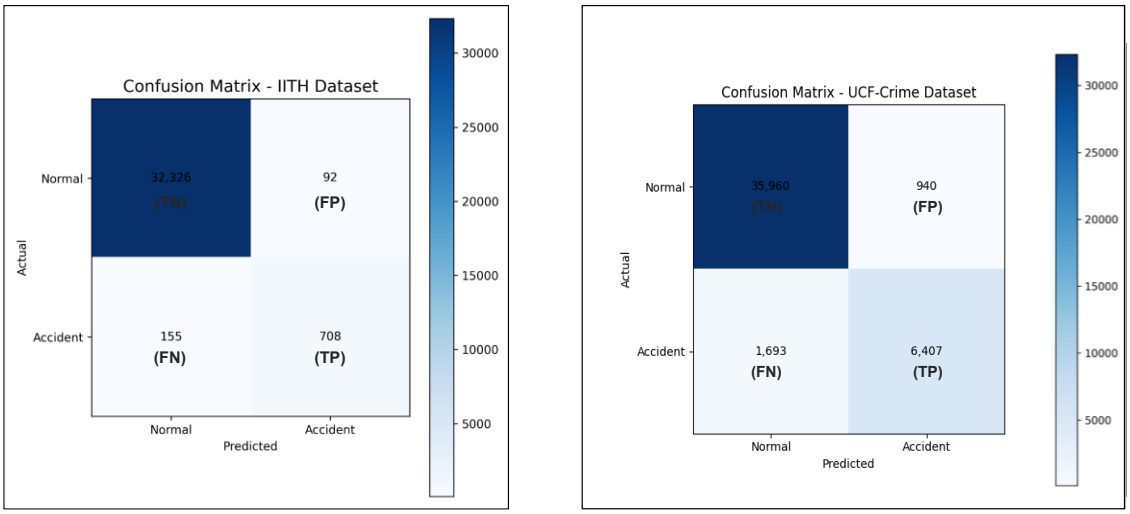
Figure 5.9: ROC curves analysis of Conv-BiLSTM Model: (a) the IITH sand (b) UCF-Crime Dataset

sets, which include 863 anomalous frames and 32,417 normal frames for the IITH dataset and 8,100 anomalous frames and 36,900 normal frames for the UCF-Crime dataset.

The CVAE-ADS model (as shown in Figure 5.10) is able to correctly classify a lot of normal frames with high true negative counts (32,326 frames for the IITH dataset, and 35,960 frames for the UCF-Crime dataset). Nonetheless, there are relatively more false negatives, 155 in IITH and 1,693 in UCF-Crime. This means that while the model is fairly successful in learning the normal traffic scenarios, there are cases where it fails to notice some of the accidents, especially those that have less distinguishable features.

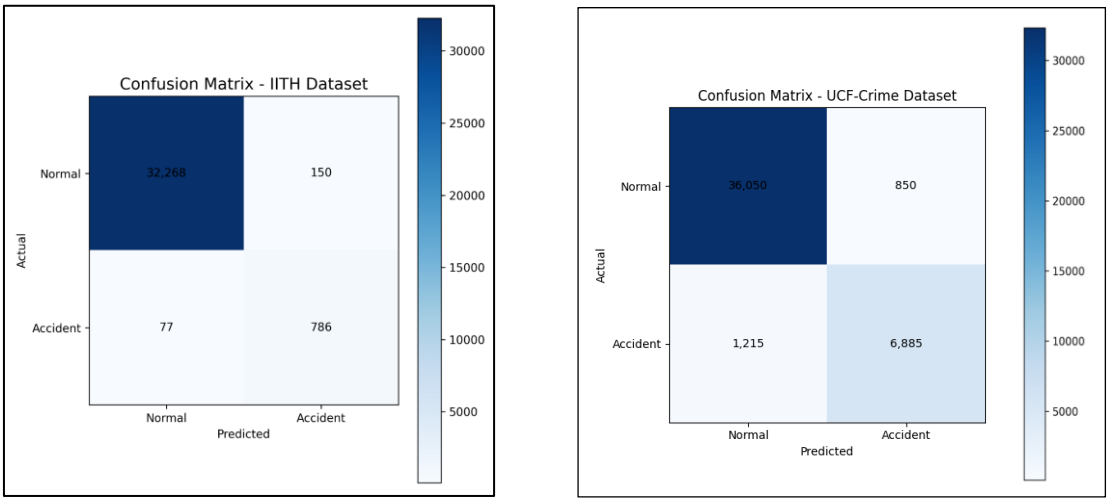
The Conv-BiLSTM model in Figure 5.11, on the other hand, shows better detection capability. The number of correctly identified accident frames becomes 786 and 6,885, and the number of missed detections turns into 77 and 1,215 for the IITH and UCF-Crime datasets, respectively. This decrease in false negatives represents a greater capability in the event of capturing accidents. While there are more false positive cases (150 in the IITH

dataset and 850 in the UCF-Crime dataset), the benefits of detecting actual accident cases are greater.



(a) **(b)**

Figure 5.10: Confusion Matrix for the proposed CVAE-ADS on (a) IITH and (b) UCF-Crime.



(a) **(b)**

Figure 5.11: Confusion Matrix for the proposed Conv-BiLSTM on (a) IITH and (b) UCF-Crime.

From both datasets, it can be seen that the Conv-BiLSTM model is able to provide a better combination of detection and false alarm rates. Missed accident events are especially significant in safety-critical applications where missing an anomaly may have serious consequences. This is consistent with the overall quantitative results and demonstrates that temporal information can be used to greatly improve the robustness and reliability of accident detection under real-world surveillance scenarios.

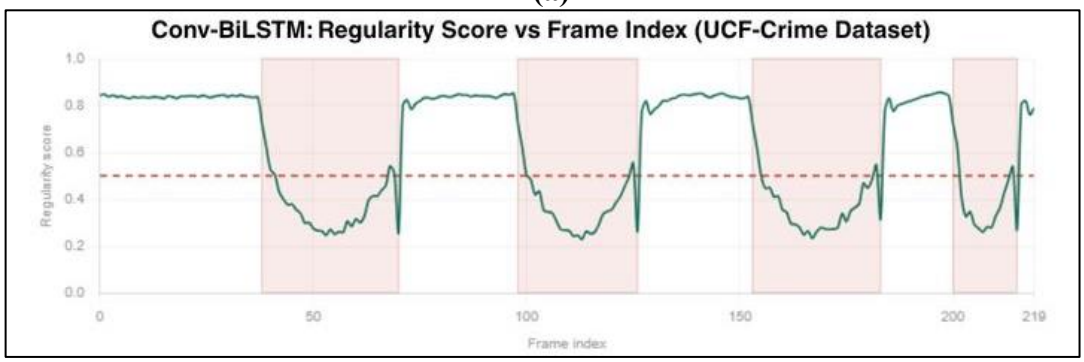
- **Regularity Score Analysis (or Temporal Analysis)**

Figure 5.12 shows the regularity score for each frame of IITH and UCF-Crime. The regularity score is a measure of how close the frame is to the learned normal frames, with higher scores corresponding to normal behaviour and lower scores corresponding to possible anomalies.

For the IITH dataset, as seen in Figure 5.12 (a), the regularity score is a high value in the normal parts of the network, showing a consistent reconstruction of the normal traffic pattern. Sharp decreases in the score are seen during accident events, with clear evidence of anomalies. Likewise, for the UCF-Crime dataset shown in Figure 5.12 (b), drops in the regularity score are associated with anomalous intervals. The differences are slightly less smooth because there is more detail in the different scenes, but the differences between normal and abnormal segments still stand out.

This threshold line tells the difference between the normal and abnormal frames, and most of the abnormal regions are below the threshold line, which indicates that the abnormal frames are detected. In general, the results demonstrate that the Conv-BiLSTM model can

(a)



(b)

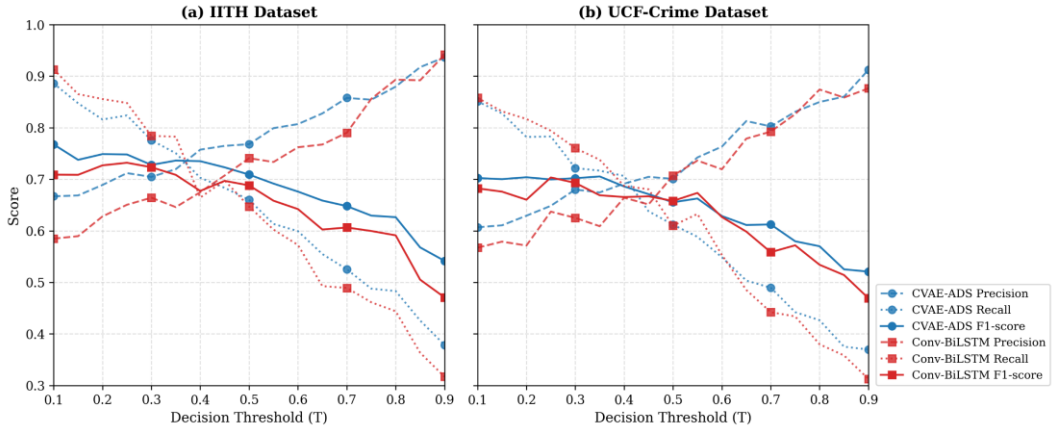
Figure 5.12: Regularity Score Vs. Frame Indexed for the proposed Conv-BiLSTM on (a) the IITH and (b) the UCF-Crime.

effectively capture the temporal changes and accurately detect anomalies in the temporal dimension.

- **Threshold Analysis**

Figure 5.13 illustrates the precision, recall, and F1 score as they change with the decision threshold for both models. A higher threshold will result in higher precision but lower recall, setting up better control over FPs and FNs.

The CVAE-ADS model performs best at moderate thresholds and has a steeper decrease in recall at higher thresholds. The F1-scores for the Conv-BiLSTM model, on the other hand,



(a) **(b)**

Figure 5.13: Precision, Recall, and F1-score vs. decision threshold (T) for CVAE-ADS and Conv-BiLSTM on (a) IITH and (b) UCF-Crime datasets

are relatively stable with precision and recall when varying the threshold across a broader range. This means increased robustness and less dependence on the choice of threshold. The best operating point is the one that falls around the middle of the threshold, maximizing the F1 score.

- **Precision–Recall Analysis:**

The precision–recall performance curves of both models are depicted in Figure 5.14. The precision of both approaches gradually declines with an increase in recall. The Conv-BiLSTM model, however, shows consistently higher precision in the majority of recall values as compared to CVAE-ADS.

The Conv-BiLSTM curve shows high dominance, which means it is able to discriminate between normal and abnormal frames well, possibly because of its capability to capture

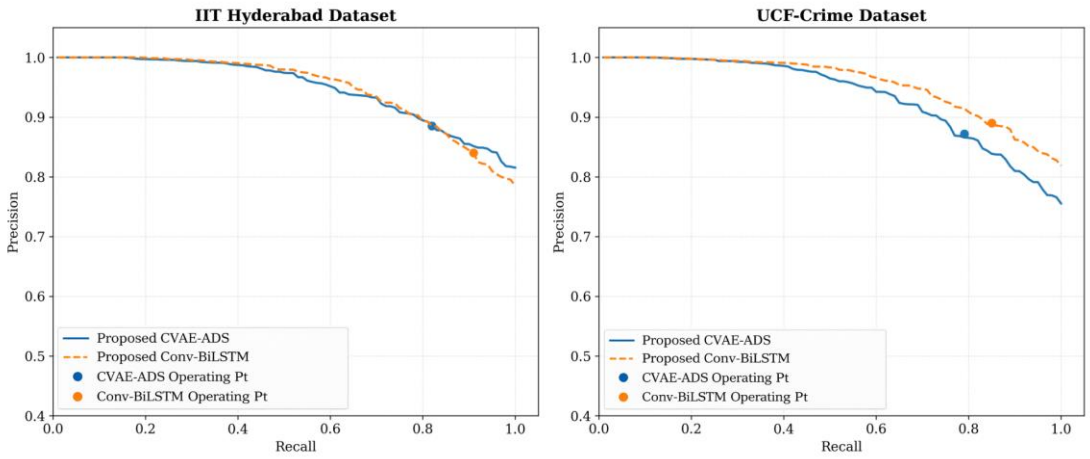


Figure 5.14: Precision–Recall (PR) curves of the proposed CVAE-ADS and Conv-BiLSTM models on (a) the IITH and (b) the UCF-Crime dataset.

temporal relations. The chosen operating points additionally demonstrate that Conv-BiLSTM is able to strike a perfect balance between precision and recall. In summary, the results show that Conv-BiLSTM is more reliable and robust for anomaly detection.

5.4.2 Video Summarization Results

The performance of the proposed models for the generation of compact and informative video summaries for both datasets is evaluated here. The evaluation is based on the important measures like reduction rate, diversity rate based on one minus SSIM, Silhouette score, and Peak signal to noise ratio.

- **Performance of CVAE-ADS Model**

The proposed CVAE-ADS model in the IITH dataset provides a high reduction rate (70% to 85%), and in the UCF-Crime dataset, it demonstrates a high reduction rate (70% to 80%), as

Table 5.4: 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

shown in Table 5.4 and Figure 5.15. This means that it works well to eliminate repeated frames and keep valuable content.

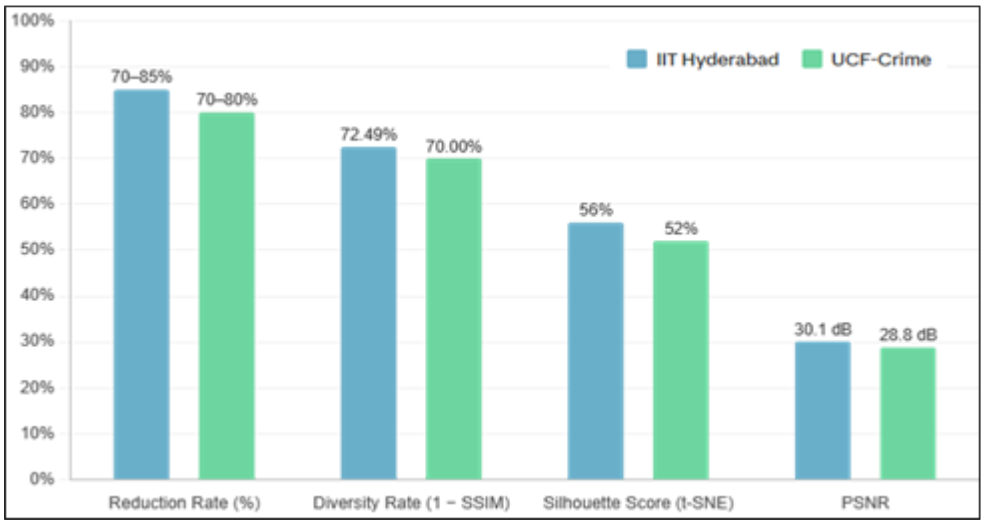


Figure 5.15: Performance Comparison of CVAE-ADS Summarization

The diversity scores for the IITH dataset and UCF-Crime dataset are 0.7249 and 0.70, respectively, indicating that the selected frames have sufficient diversity. Moreover, the silhouette scores are 0.56 and 0.52, which shows that the latent representations are reasonable. clustered. The PSNR values of 30.1 dB and 28.8 dB also verify that the reconstructed summaries are of acceptable visual quality.

Since it is important to assess the practical effectiveness of the proposed approaches, sample-based evaluation is performed for some of the selected videos from both data sets. An important condition to note is that only a portion of the videos is shown, and this analysis is therefore indicative and not a full comparison.

The performance of the CVAE-ADS approach on sample IITH and UCF-Crime videos is shown in Tables 5.5 and 5.6, respectively. The model can summarize the IITH dataset with a 71%–77% reduction while keeping most accident frames. For instance, in Video 1, 240 frames out of 305 are kept with reduction 77.1% and compression 4.20×. For the UCF-Crime dataset, similar trends are observed, and the amount of reduction is between 73% and 79%. For example, Video 2 has compression of 4.68× and a reduction of 78.6% and retains 278 frames. Overall, CVAE-ADS is well balanced between accident frame retention and compression, but a few frames are lost.

Table 5.5: 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 5.6: 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

- **Performance of Conv-BiLSTM Model**

The performance of the summarization approach based on Conv-BiLSTM is given in Table 5.7 and Figure 5.16 below. The model provides a similar reduction range (65%-85%) and shows better diversity (0.8567 and 0.82) and silhouette scores (0.64 and 0.55) on both sets of data.

Moreover, the PSNR is 32.45 dB and 31.3 dB, which are higher than the CVAE-ADS model, suggesting that the reconstruction quality is better and visual details are preserved. This indicates that the Conv-BiLSTM model has more capability to capture the key temporal and structural information when summarizing.

Table 5.7: 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

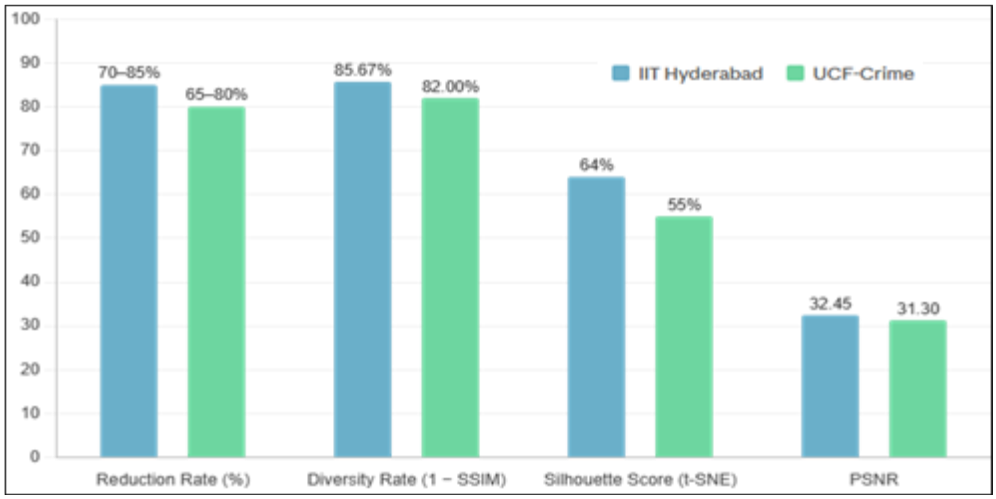


Figure 5.16: Performance Comparison of Conv-BiLSTM

Tables 5.8 and 5.9 give a summary of the CONV-BiLSTM model performance. For IITH data, the model has a better effect in reducing the data (77%-82%) and a better retention effect. For instance, video 1 has 250 frames, 80.3% reduction, and 5.08 compression. For the UCF-Crime dataset, the reduction ranges from 77 to 81, and the data is similar across videos. For example, in the case of Video 5, the compression ratio is 4.15 x and the reduction is 80.9%. Overall, CONV-BiLSTM presents a larger reduction and maximum compression with competitive accent frames as compared to CVAE-ADS.

Table 5.8: Performance Evaluation on Sample IITH Videos

Video	Original Frames	Accident Frames	Retained Accident (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 5.9: Performance Evaluation on Sample UCF-Crime Videos

Video	Original Frames	Accident Frames	Retained Accident (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

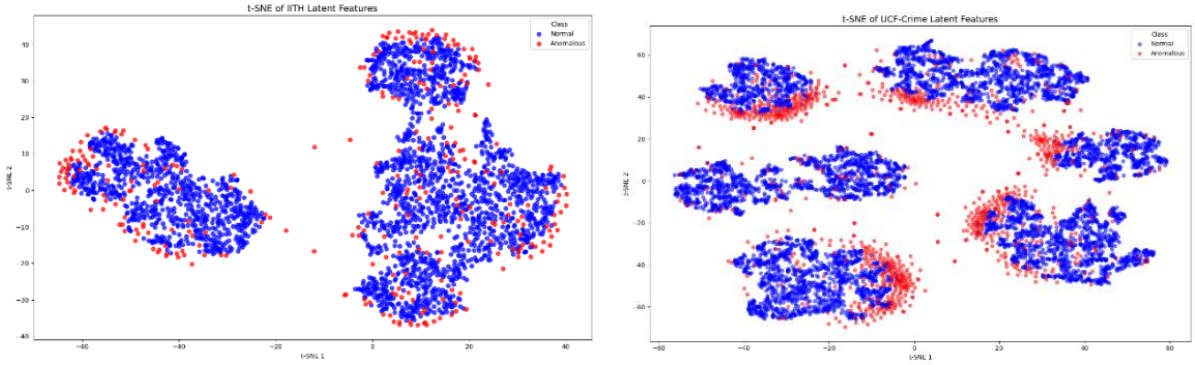
5.5 Qualitative Analysis

In this section, the proposed models will be qualitatively assessed; this will be done by visual analysis in addition to the quantitative analysis. Evaluation is done based on the separability of the features, the quality of the reconstruction, and the relevance of the generated summary.

5.5.1 Latent Space Visualization

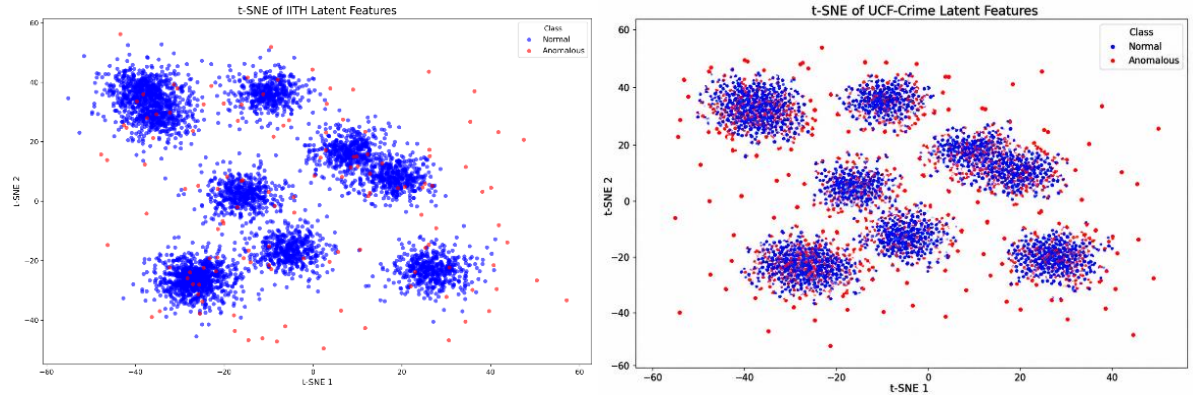
Figures 5.17 and 5.18 show the latent feature visualization using t-SNE learnt by the CVAE-ADS model for both the datasets compared with the Conv-BiLSTM model. Figure 5.17 demonstrates that the CVAE-ADS model exhibits some separation between the normal and anomalous frames, but there is still overlap, suggesting the limited discriminative property of the latent space. By contrast, as shown in Figure 5.18, the Conv-BiLSTM model can create tighter and more coherent clusters and provides a separation between normal and anomalous instances, especially for the IITH data set. This implies that temporal modelling can better represent features by capturing sequential dependency.

There is still some overlap in the UCF-Crime dataset, but the cluster structure is more organized when compared to CVAE-ADS, as the scenes are more complex. The denser intra-class grouping and the distinct inter-class boundaries indicate good feature discrimination.



(a) **(b)**

Figure 5.17: t-SNE of Latent Features for CVAE-ADS Model on (a) IITH and (b) UCF-Crime



(a) - **(b)**

Figure 5.18: t-SNE of Latent Features for Conv-BiLSTM Model on (a) IITH and (b) UCF Crime

These observations validate that Conv-BiLSTM produces more structured and separable latent representations, which yield better anomaly detection performance.

5.5.2 Reconstruction Quality Analysis

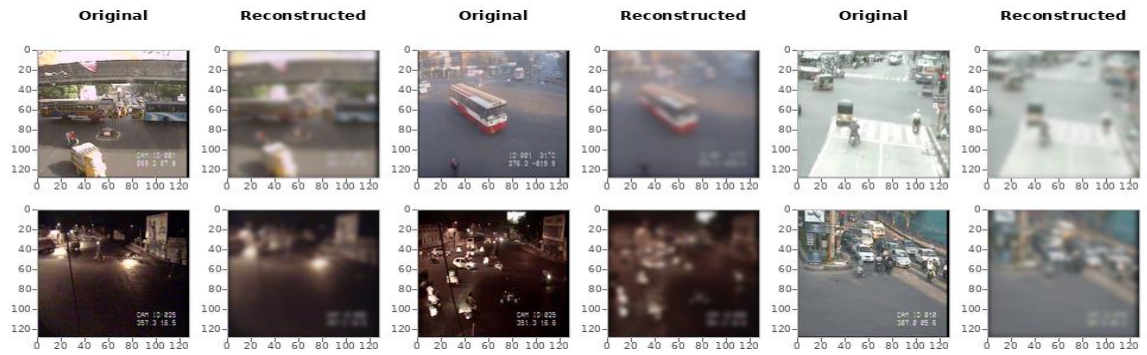
The visual comparison of the original and reconstructed frames as shown in Figure 5.19 and Figure 5.20 shows the reconstruction capability of both the models for IITH accident and UCF-Crime datasets. In regular traffic situations, the reconstructed frames are very similar to the original inputs, showing that the models are able to well capture and preserve the underlying spatial characteristics.

Contrarily, frames with abnormal events exhibit significant blur and loss of fine details in the reconstructed outputs. This is to be expected in reconstruction-based approaches because the models are trained mostly on normal patterns and are not able to reconstruct the unseen anomalies accurately.

(a)

(b)

Figure 5.19: Visual results of original and reconstructed frames for (a) IITH and (b) UCF-Crime generated by the CVAE-ADS



(a)

(b)

Figure 5.20: Visual results of original and reconstructed frames for (a) IITH and (b) UCF-Crime generated by the Conv-BiLSTM model

The comparison shows that the Conv-BiLSTM model generates more structurally consistent and well-represented objects than CVAE-ADS. The quality of reconstruction has been improved with better-preserved spatial details and less distortion, which indicates good temporal modelling. In general, the visual results confirm the finding that reconstruction errors are more significant in areas of anomalies and thus are good indicators for anomaly detection.

5.5.3 Summary Quality Assessment

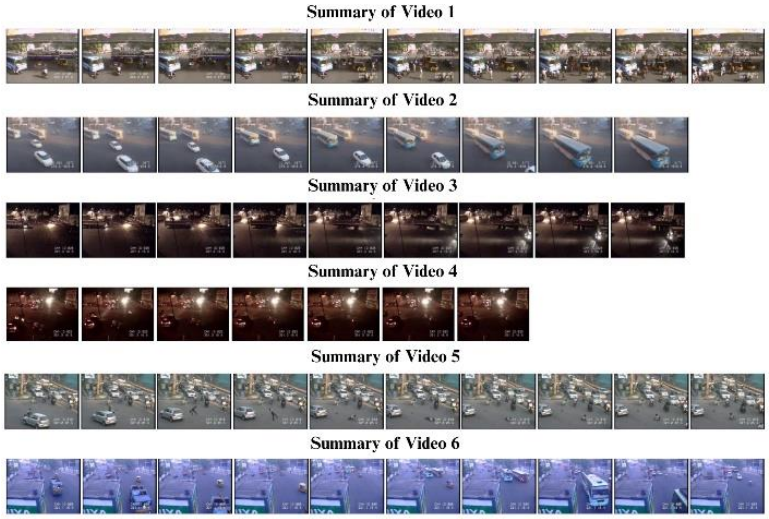
In Figure 5.21 and Figure 5.22, the qualitative summarization of the sample videos of the IITH and UCF-Crime datasets by both proposed models is discussed. It should be pointed out that there are a few video clips shown to be visually illustrated.

The CVAE-ADS model demonstrates that the model is able to capture important moments related to accidents and eliminate redundant frames. The resulting summaries are brief and focus on major events, with occasional inclusion of some irrelevant frames indicating a fairly cautious summarization approach. The summaries generated by the model Conv-BiLSTM are more accurate and detailed, with the frames identified being more localized to the occurrence of the accident. The model generates fewer and more accurate summaries, minimizing redundancy but preserving significant information on time.

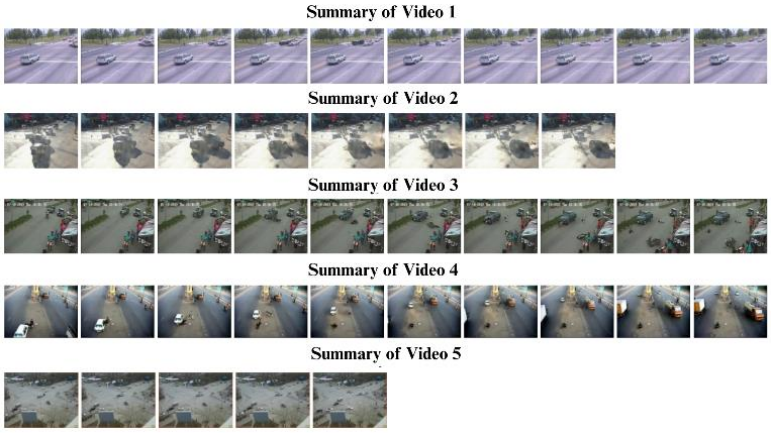
Comprehensively, the two models have effective balances between compression and information retention. Though CVAE-ADS is more efficient for summarization with less complexity, Conv-BiLSTM can provide more diversity, quality of clustering, and reconstruction. This highlights the advantage of temporal modelling to create more useful accident segments. However, the CONV-BiLSTM generated quite concise, informative, and credible video summaries. Both models have proven success in the task of outlining narrower summaries, and CVAE-ADS provides both consistent and stable coverage of events.

5.6 Comparative Analysis with State-of-the-Art Methods

The comparative research of the current state-of-the-art techniques for accident detection available in the literature is presented in Table 5.10, which helps to prove the effectiveness of the suggested techniques in different benchmarks. Previous studies like Singh et al. [119] and Pawar et al. [132] have utilized unsupervised techniques of autoencoders, but with considerably lower AUC (77% and 79%) and larger numbers of EER for the IITH dataset.

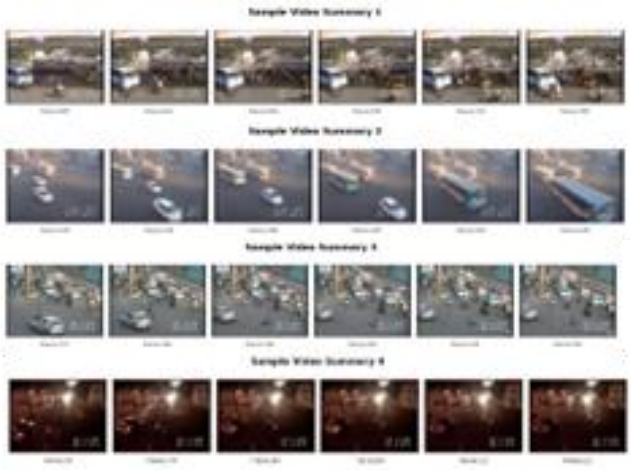


(a)

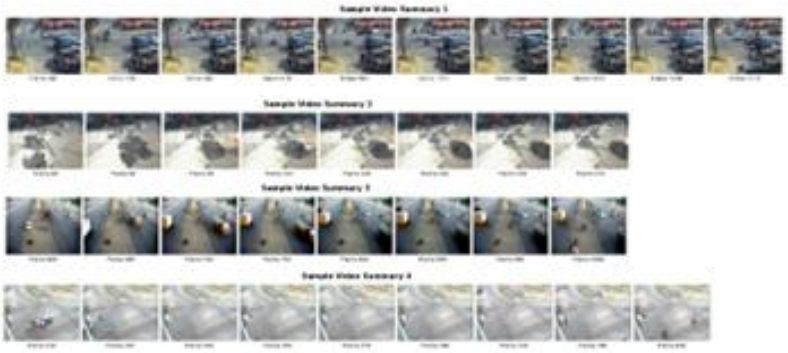


(b)

Figure 5.21: Example outcomes of video summarization by CVAE-ADS: (a) IITH and (b) UCF-Crime.



(a)



(b)

Figure 5.22: Example outcomes of video summarization by Conv-BiLSTM: (a) IITH, (b) UCF-Crime

Supervised algorithms are based on complicated structures and labelled data, and they obtain better results in detection (up to 92.5%) and AUC (96.32%) [130]. 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 AUCs of 94.60% (IITH) and 91.80% (UCF-Crime) and lower EERs (14.97% and 16.20%), which implies a better detection rate and strength. Overall, the proposed models offer a more balanced performance and practicality compared to existing models, especially in terms of dependence on labelled data while achieving high detection accuracy.

Table 5.10: Comparison with State-of-the-Art Accident Detection.

Ref.	Learning	Approach	Dataset	AUC	EER	mAP	Detection Rate	FAR
[119]	Unsupervised	Unsupervised Denoising Autoencoder + One-Class SVM	IITH	77%	22.50 %	-	-	-
[129]	Supervised	DETR + Random Forest		82%	-	-	78.2%	-
[130]	Supervised	Retinex + YOLOv3 + Decision Tree	Online CCTV videos	96.32%	-	-	92.5%	7.5 %

[131]	Supervised	CNN + Rolling Prediction	VAID & Test Dataset	-	-	-	88%	-
[132]	Unsupervised	Conv-AE + Seq2Seq LSTM Autoencoder	IITH	79%	20.50 %	60%	-	-
			DOTA	84.70%	11.7 %	-	-	-
[133]	Supervised	YOLOv2 + Transfer Learning	IITH	-	-	76%	-	-
[134]	Semi-Supervised	Object Interaction + Refinement + Heatmaps	UCF Crime	69.70%	-	-	-	0.8
			CADP	72.59%	-	-	-	2.2
[135]	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 %			

As shown in Table 5.11, the comparison with the already available video summarization methods demonstrates that the proposed methods are quite efficient in reducing the diversity and being visually appealing. Thomas et al. [136] and Kosambia et al. [137] have proposed some algorithms in the past, which are primarily perceptual or Spatio-temporal, but don't give quantitative values like reduction rate, diversity, PSNR, etc. There are also some algorithms that have been proposed by Pramanik et al. [138]. Moderate reduction rates (around 20-50 %) are reported in more recent methods such as YOLOv5 [139] and Saxena et al. [140], which don't provide diversity assessment and quality evaluation. Conversely, the CVAE-ADS solution is much more effective in reducing the frame (70-85% and 70-

80%, respectively) and retains good diversity (0.7249 and 0.70) and reasonable visual quality (PSNR up to 30.1 dB). Moreover, the results show the effectiveness of the proposed Conv-BiLSTM method with high reduction rate (up to 85%), high diversity (up to 0.8567), and high reconstruction quality (PSNR up to 32.45 dB). The proposed models are, in general, better than the existing ones, providing a more balanced compromise between compression, diversity, and visual quality and are therefore more suited for practical application in video summarization.

Table 5.11: Comparison of state-of-the-art video summarization techniques

Ref.	Approach / Model	Dataset(s) Used	Reduction Rate (%)	Diversity Rate	PSNR (dB)
[136]	Perceptual Video Summarization	YouTube-8M, Urban Tracker	Used saliency	×	×
[137]	Video Synopsis using ResNet	IITH	×	×	×
[138]	Z-STRFG: Spatio-temporal fuzzy approach	YouTube8M, Anomaly20, Urban Tracker	×	×	×
[139]	YOLOv5 + Privacy-preserved summarization	Synthetic + Real-time Data	42.97%	×	×
[140]	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%	0.8567	32.45
		UCF-Crime	65–80%	0.82	31.3

5.7 Chapter Summary

In this chapter, a complete experimental evaluation of the proposed event-based video summarization system for the traffic surveillance application was presented. The evaluation was performed on two publicly available benchmark data sets, the IITH data set and the UCF-Crime (accident subset) under a common unsupervised learning scenario with both models trained only with normal traffic data. To evaluate the accident detection and video summarization component of the framework, a set of well-defined performance metrics were used: AUC, EER, reduction rate, diversity score, and PSNR.

The proposed CVAE-ADS model was shown to be reliable on both datasets with an AUC of 0.912 and 0.893 and corresponding EER of 16.5% and 18.7% on IITH and UCF-Crime, respectively. On the summarization side, it achieved a considerable compression ratio of 70-85% with a diversity score of 0.7249 and PSNR of 30.1 dB on IITH. The results show that the spatial probabilistic latent representation learnt by CVAE-ADS is useful for normal traffic modelling and for providing compressed informative event summaries without relying on any labelled data.

Conv-BiLSTM Autoencoder was superior to CVAE-ADS in all aspects of the evaluation: 0.9460 AUC on IITH and 0.918% AUC on UCF-Crime, and lower EER of 14.97% and 16.20%, respectively. It had equally good summarization results with a maximum reduction of up to 85%, a diversity score of 0.8567, and a PSNR of 32.45 dB. Such improvements are due to the benefits of bidirectional sequence learning, which has the capability of modelling both spatial appearance and temporal dynamics in order to better capture motion-based anomalies as well as temporally coherent event structures compared to frame-level methods.

Both proposed models achieved competitive and, in some cases, better performance than the current state-of-the-art methods, especially because they are completely unsupervised. Although supervised methods, like Retinex + YOLOv3 or DETR + Random Forest, with annotated accident samples and complex multi-stage pipelines, can obtain higher detection rates, the proposed models can still obtain similar or higher detection results without relying on annotated accident samples. For video summarization, the presented methods achieved significant improvement over previous video summarization methods in terms of compression efficiency and visual quality, further supporting the feasibility of latent space-based summarization.

In this chapter, the results from the findings of the chapter have been presented, that proves the proposed intelligent traffic surveillance framework to be effective and feasible. In the next chapter, the overall conclusions of this research are presented, the limitations of this research are discussed, and the directions for future research are highlighted.

Chapter 6: Conclusion and Future Scope

6.1 Summary of Research Work

As the number of smart city infrastructure continues to grow, the amount of traffic surveillance footage has also increased substantially, making it increasingly impractical to manually monitor all such footage to detect accidents and prove events in real-time. The study aimed to fill this need with an automated, intelligent, and annotation-free system to detect accident events from continuous traffic surveillance streams and summarize them into short and informative video snippets.

In this regard, an end-to-end event-based video summarization framework was proposed and tested, based on two complementary deep learning architectures. The first one is the Convolutional Variational Autoencoder for Accident Detection and Summarization (CVAE-ADS), which is designed to learn probabilistic latent representations of normal traffic frames. The model estimates the spatial distribution of the regular traffic patterns and flags as anomalies the frames whose reconstruction quality falls below a threshold, which is computed from the statistics of the model. The second architecture was the Conv-BiLSTM Autoencoder, which built on this by introducing bidirectional temporal models using stacked BiLSTM layers in the autoencoder on sliding window frame sequences. The combination of spatial and temporal learning enabled the model to learn the continuity of movements, abrupt changes, and transitions between events, which are essential for accurate accident detection in dynamic surveillance settings.

Both models were trained using only normal traffic data in an unsupervised learning paradigm to avoid the need for expensive and scarce labelled accident data. Both architectures were then subjected to a common summarization pipeline, which was based on a video. To encode the anomalous frames into latent representations, cluster them by the K-Means algorithm based on the visual similarity and represent them by centroid-based selection by reducing them to keyframes. Then, using SSIM, only those frames that were visually different were kept, and the keyframes were sorted by time to create a short summary video focused on events. The entire framework has been tested on the IITH and UCF-Crime benchmark datasets, and the results show good performance in accident detection and summarization.

6.2 Key Contributions

The key findings from this research are as follows:

- **Design of the CVAE-ADS Framework:** A novel framework of convolutional variational autoencoder (CVAE) for joint anomaly detection and summarization for smart traffic monitoring was proposed. The model can learn the normal traffic distribution behaviour, as well as capture deviations from this distribution as indications of accident events, without the need for any labelled anomaly data, by combining linear convolutional feature extraction with a probabilistic latent space.
- **Integration of Conv-BiLSTM for Spatio-Temporal Anomaly Detection:** The Conv-BiLSTM autoencoder was designed to model the spatial and temporal characteristics of traffic video sequences.

  The two bidirectional LSTM layers (over the sequences of the sliding window) enable the model to learn both forward and backward temporal relationships, resulting in enhanced detection accuracy and robustness of motion-based anomalies compared to frame-based approaches.
- **Adaptive Thresholding Strategies:** A statistical threshold defined by the mean and standard deviation of the reconstruction errors for the CVAE-ADS and a threshold defined by Median Absolute Deviation (MAD) and Exponential Moving Average (EMA) smoothing for the Conv-BiLSTM model. These adaptive strategies increase detection stability and decrease noise and environmental variability sensitivity.
- **Unified Latent-Space-Driven Summarization Pipeline:** A summarization framework that can be used with both proposed models was developed, instead of using the pixel-level approach. Combining K-Means clustering, centroid-based keyframe selection, and SSIM-based redundancy removal allows generating semantically meaningful and compact video summaries that are able to accurately capture the progression of the accident event.
- **Annotation-Free Unsupervised Learning Paradigm:** Both proposed models are trained with normal traffic footage and don't require any labelled training data. This renders the framework applicable and efficient in surveillance scenarios with little or no annotated accident data.
- **Comprehensive Benchmarking on Public Datasets:** The framework was extensively tested over two publicly available datasets (IITH and UCF-Crime), and it outperforms

or matches the current state-of-the-art supervised and unsupervised methods with significantly reduced need for annotated data.

6.3 Limitations of the Proposed Framework

However, the results obtained using the proposed framework were encouraging with the CVAE-ADS and Conv-BiLSTM models achieving 92.5% and 94.0% accuracy on the IITH accident dataset and 88.6% and 90.0% on the UCF-Crime accident subset, with AUC values of up to 0.946, which should be considered in the context of the experimental setting. The limitations of contributions made by this research are therefore as follows:

First, the experimental validation was limited to two benchmark datasets: (a) the IIT Hyderabad (IITH) accident dataset, captured from the fixed CCTV surveillance network of Hyderabad City, and (b) a road-accident subset of the UCF-Crime dataset, which is real-world surveillance footage captured from low-resolution cameras. These datasets, collectively, have a limited range of traffic scenarios and camera viewpoints that are useful. The numbers presented are for those configurations; it has yet to be proven if similar results can be achieved on bigger surveillance networks or in other geographic and infrastructural scenarios.

Second, the models developed in this work are reconstruction based, since both datasets consist of fixed-angle CCTV footage; robustness to camera motion, pan–tilt–zoom operation and aerial viewpoints was therefore not evaluated. Although the footage naturally spans diverse conditions like daytime, nighttime, and early-morning lighting, varying weather, traffic congestion, and partial occlusions. In the data, such as heavy fog, storms, or severe occlusion, therefore, remains untested, and no claim of robustness under such conditions is made. As presented in this thesis, the adaptive thresholding mechanism increases tolerance to slower scene changes but would require specific testing on condition-annotated data to be declared reliable in very dynamic scenarios.

Third, all experiments were performed on workstations with GPUs. The Conv-BiLSTM model, in particular, due to its temporal sequence processing, has a computational cost that has not been optimized for low-power hardware. The deployment in real time on edge/embedded surveillance devices was not considered within the scope of this work.

Lastly, a framework was designed and evaluated for the identification of single accident events. It was not tested in multiple events or events that happened close to each other in time in dense traffic scenes.

6.4 Conclusion

This research focused on the need for intelligent and automated traffic surveillance systems that can efficiently analyse and process the vast amount of continuous video data. Conventionally deployed surveillance techniques include manual inspection or supervised learning methods, which need a vast amount of labelled data. To address these drawbacks, this work presented a unified event-based video summarization framework for traffic surveillance through a combination of anomaly detection and automatic video summarization. The objective was to identify such anomalies in the framework and produce concise summaries that still carry a lot of useful information from the continuous surveillance footage, without any labelled data of accidents.

Two deep learning architectures, CVAE-ADS and Conv-BiLSTM Autoencoder, were developed and evaluated. In the CVAE-ADS framework, probabilistic latent representation learning was used to model normal traffic behaviour, and anomalies were detected using the reconstruction error analysis. The Conv-BiLSTM approach, on the other hand, combined spatial feature extraction by using a convolutional network with temporal sequence modelling by the BiLSTM network, aiming to model traffic dynamics from both appearance and motion perspectives. The use of adaptive thresholding mechanisms, latent-space clustering, centroid-based keyframe selection, and redundancy removal based on SSIM allowed the framework to produce short and event-centric summaries, which do not lose any important information about accidents.

These proposed approaches were thoroughly tested on both the IITH and UCF-Crime datasets with quantitative and qualitative performance measures. The experimental results showed that the proposed models had high anomaly detection ability and good summarization performance in reality traffic surveillance environment. The CVAE-ADS model successfully reached the detection accuracy with computational efficiency, making it appropriate for resource-aware surveillance environments. The Conv-BiLSTM model consistently performed best because it can model temporal dependencies and motion continuity from the video sequences. Additionally, the compression efficiency, diversity,

and visual quality of the generated summaries were very high, which verified the effectiveness of the latent-space-driven summarization framework.

The proposed framework was also compared with the state-of-the-art techniques, and the comparative analysis further proved the effectiveness of the proposed framework. Also, the proposed models do not require complex multi-stage processing pipelines or annotated accident data like many supervised approaches. This clearly shows the importance of reconstruction-based learning in intelligent surveillance systems, especially in real-world applications with restricted, costly, or inaccessible anomaly data. The combined anomaly detection and summarization within a single framework also enhances the efficiency of the operations by both cutting down on storage space needed and manual review time.

In sum, the results of this study show that reconstruction, unsupervised deep learning, and latent space-driven summarization can be an effective and efficient method for intelligent traffic surveillance. The proposed framework also helps in the development of automated systems that can assist in the monitoring of traffic, the analysis of accidents, emergency response, and smart city management. This research combines both anomaly detection and event summarization into a single framework, which has the potential to provide a solid base for future development of intelligent video surveillance systems and traffic analysis systems.

6.5 Future Scope

The limitations described in Section 6.3 indicate several specific directions in which this work can be pursued.

Although the proposed models performed well across the diverse conditions present in the evaluation data, such as daytime, night-time, and early-morning lighting, varying weather, traffic congestion, and partial occlusions, their behaviour under extreme scenarios that are absent or rare in the data, such as heavy fog, storms, or severe occlusion, remains untested. Future work should therefore evaluate the models on larger surveillance collections that explicitly include such extreme conditions, so that robustness can be quantified for each scenario individually. Extending the evaluation to moving or pan–tilt–zoom cameras would further relax the fixed-camera assumption underlying the present reconstruction-based models. Transformer architectures and attention mechanisms can be added to the Conv-BiLSTM pipeline to better capture long-range temporal dependencies, while multi-camera

fusion could alleviate the dependence on a single viewpoint and make detection more reliable for occlusion.

In order to overcome the above-mentioned computational limitations, further studies will be conducted on model compression techniques, including pruning, quantization, and knowledge distillation, and their application to the model of this thesis to ensure the detection accuracy obtained in this research is maintained on resource-limited edge and embedded platforms. This would move the framework from an offline evaluation to practical real-time deployment in operational surveillance networks.

Single event assumption is also worth considering. To expand the detection stage so that it could detect multiple accidents happening in real time, or in quick succession in dense traffic, would necessitate a rethinking of the anomaly scoring as well as the event-segmentation logic, and would be a logical progression towards more complex deployments in urban settings. In addition to visual data, incorporating complementary modalities like audio signals, roadside sensors, or vehicle telemetry could help to disambiguate events that are similar in appearance and enhance decision support in such cases.

Lastly, the event-based summarization module created in this thesis could be extended to include semantic-aware approaches that would add temporal labels and automatically generated descriptions to the events detected. These sorts of summaries would be more directly usable by traffic authorities and emergency response teams, making the overall system more interpretive.

Collectively, these directions further the contributions made in this research and provide a roadmap for future efforts to create more general, efficient, and interpretable traffic surveillance systems.

References

[1] A. Noor, H. Almukhalfi, T. H. Noor, and R. Ranjan, "TAERM: Traffic Accident Emergency Response Management Framework for Detection and Classification Using IoT and YOLOv9," *Future Generation Computer Systems*, p. 108444, 2026.

[2] X. Li, X. Ye, and Q. Sun, "Leveraging Connected Vehicle Data for Near-Crash Detection and Analysis in Urban Environments," *arXiv preprint* arXiv:2409.11341, 2024.

[3] Bazzan, A.L.C. Artificial Intelligence in Transportation: A Meta Review. SN COMPUT. SCI. 7, 172 (2026). https://doi.org/10.1007/s42979-025-04530-z

[4] S. B. Sabale and V. N. Khedkar, "A comprehensive survey of transformer-based models for video anomaly detection in surveillance systems," *International Journal of Safety and Security Engineering*, vol. 15, no. 10, pp. 2143–2154, 2025, doi: 10.18280/ijsse.151017.

[5] A. Khajuria, A. Kaul, and A. Mahajan, "A Review of Deep Learning Techniques for Smart Surveillance and Public Safety," in *Intelligent Vision and Computing. ICIVC 2025. Lecture Notes in Networks and Systems*, vol. 1711, A. K. Saha, H. Sharma, M. Prasad, and P. K. Gupta, Eds. Springer, Cham, 2026, doi: 10.1007/978-3-032-10667-4_19.

[6] U. U. Deshpande, S. Shanbhag, R. Patil, R. A. A. Chate, S. Das Chagas Silva Araujo, K. Pinto *et al.*, "Automatic two-wheeler rider identification and triple-riding detection in surveillance systems using deep-learning models," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 104, 2025.

[7] O. Alzamzami, Z. Alsaggaf, R. AlMalki, R. Alghamdi, A. Babour, and L. Al Khuzayem, "Passable: An Intelligent Traffic Light System with Integrated Incident Detection and Vehicle Alerting," *Sensors*, vol. 25, no. 18, p. 5760, 2025.

[8] W. Sun, L. N. Abdullah, F. B. Khalid, and P. Suhaiza Binti Sulaiman, "EAR-CCPM-Net: A Cross-Modal Collaborative Perception Network for Early Accident Risk Prediction," *Applied Sciences*, vol. 15, no. 17, p. 9299, 2025.

[9] P. Chaware, V. Dhamdhere, M. Sawane, S. Patil, P. Sarode, and K. Ukey, "Development of Intelligent Video Surveillance System Using Deep Learning and

Convolutional Neural Networks: A Proactive Security Solution," *Engineering Proceedings*, vol. 114, no. 1, p. 11, 2025, doi: 10.3390/engproc2025114011.

[10] M. K. Bhosale, S. B. Patil, and B. B. Patil, "Automatic Video Traffic Surveillance System with Number Plate Character Recognition Using Hybrid Optimization-Based YOLOv3 and Improved CNN," *International Journal of Image and Graphics*, vol. 25, no. 04, p. 2550041, 2025.

[11] A. Ismail, A. R. Ismail, N. A. Shaharuddin, M. A. Ara, A. A. Puzi, S. Awang, and R. Ramli, "Vision-based vehicle classification for smart city," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 2, pp. 441–453, 2025.

[12] M. S. Ghauri, U. I. Bajwa, G. Saleem *et al.*, "SurveillanceNet: Spatio-temporal anomaly identification in surveillance videos using two-stream CNN and LSTM," *Multimed Tools Appl*, vol. 84, pp. 45871–45895, 2025, doi: 10.1007/s11042-025-20961-5.

[13] G. Madhumitha and R. Senthilnathan, "Development of a deep learning image classification network based on vehicular collision using instance mask guided attention," *Signal, Image and Video Processing*, vol. 19, no. 4, p. 342, 2025.

[14] M. S. Arefin, M. I. S. Mahin, and F. A. Mily, "Real-time rapid accident detection for optimizing road safety in Bangladesh," *Heliyon*, vol. 11, no. 4, 2025.

[15] K. Srinivasan, "Object tracking using optimized Dual interactive Wasserstein generative adversarial network from surveillance video," *Knowledge-Based Systems*, vol. 311, p. 113084, 2025.

[16] A. Niaz, S. Ul Amin, S. Soomro, H. Zia, and K. Nam Choi, "Spatially Aware Fusion in 3D Convolutional Autoencoders for Video Anomaly Detection," *IEEE Access*, vol. 12, pp. 104770–104784, 2024, doi: 10.1109/ACCESS.2024.3435144.

[17] Z. Xi, X. Liu, Y. Liu, Y. Cai, and Y. Zheng, "Integrating generative adversarial networks and convolutional neural networks for enhanced traffic accidents detection and analysis," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[18] Z. Shi, Y. Wang, D. Guo, F. Jiao, H. Zhang, and F. Sun, "The urban intersection accident detection method based on the GAN-XGBoost and Shapley additive explanations hybrid model," *Sustainability*, vol. 17, no. 2, p. 453, 2025.

[19] S. Jaradat, M. Elhenawy, H. I. Ashqar, A. Paz, and R. Nayak, "Leveraging deep learning and multimodal large language models for near-miss detection using crowdsourced videos," *IEEE Open J. Comput. Soc.*, vol. 6, pp. 223–235, 2025.

[20] E. Dilek and M. Dener, "Enhancement of Video Anomaly Detection Performance Using Transfer Learning and Fine-Tuning," *IEEE Access*, vol. 12, pp. 73304–73322, 2024, doi: 10.1109/ACCESS.2024.3404553.

[21] Y. Xu, H. Hu, C. Huang, Y. Nan, Y. Liu, K. Wang *et al.*, "TAD: A large-scale benchmark for traffic accidents detection from video surveillance," *IEEE Access*, vol. 13, pp. 2018–2033, 2024.

[22] S. Natha, M. Siraj, S. A. Alsaif *et al.*, "A fusion approach of YOLOv8 and CNN-Transformer for End-to-End road anomaly detection," *Sci Rep*, vol. 15, p. 45341, 2025, doi: 10.1038/s41598-025-29718-4.

[23] A. Saraff *et al.*, "Indian Traffic Surveillance Video Summarization Using YOLO and Multi-Level Masking," *IEEE Access*, vol. 13, pp. 171371–171385, 2025, doi: 10.1109/ACCESS.2025.3616267.

[24] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, and S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," *Expert Syst. Appl.*, vol. 237, PC, Mar. 2024, doi: 10.1016/j.eswa.2023.121288.

[25] Y. Xia, N. Qian, L. Guo, and Z. Cai, "CF-SOLT: Real-time and accurate traffic accident detection using correlation filter-based tracking," *Image and Vision Computing*, vol. 152, p. 105336, 2024.

[26] S. Rani and S. Dalal, "Comparative analysis of machine learning techniques for enhanced vehicle tracking and analysis," *Transportation Engineering*, vol. 18, p. 100271, 2024.

[27] A. Ahmed, M. Farhan, H. Eesaar, K. T. Chong, and H. Tayara, "From detection to action: A multimodal AI framework for traffic incident response," *Drones*, vol. 8, no. 12, p. 741, 2024.

[28] C. Liang, Q. Chen, Q. Li, Q. Wang, K. Zhao, J. Tu, and A. Jafaripournimchahi, "HADNet: A novel lightweight approach for abnormal sound detection on highway based on 1D convolutional neural network and Multi-Head Self-Attention mechanism," *Electronics*, vol. 13, no. 21, p. 4229, 2024.

[29] B. Cai, Y. Feng, X. Wang, and M. Quddus, "Highly accurate deep learning models for estimating traffic characteristics from video data," *Applied Sciences*, vol. 14, no. 19, p. 8664, 2024.

[30] J. Park, J. Lee, Y. Park, and Y. Lim, "Deep learning-based stopped vehicle detection method utilizing in-vehicle dashcams," *Electronics*, vol. 13, no. 20, p. 4097, 2024.

[31] V. Mandal, A. R. Mussah, P. Jin, and Y. Adu-Gyamfi, "Artificial intelligence-enabled traffic monitoring system," *Sustainability*, vol. 12, no. 21, p. 9177, 2020.

[32] S. S. Chintalapati, K. C. Krishna, and I. T. R. Madhav, "Analysis on IoT environment for detection and prevention of road accidents with communication modules," *Journal of Green Engineering*, vol. 10, no. 11, pp. 12096–12106, 2020.

[33] D. M. Tan and L. M. Kieu, "TRAMON: An automated traffic monitoring system for high density, mixed and lane-free traffic," *IATSS research*, vol. 47, no. 4, pp. 468–481, 2023.

[34] S. Sreedhar, A. O. Philip, and M. U. Sreeja, "Autotrack: a framework for query-based vehicle tracking and retrieval from CCTV footage using machine learning at the edge," *International Journal of Information Technology*, vol. 15, no. 7, pp. 3827–3837, 2023.

[35] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Real-time event-driven road traffic monitoring system using CCTV video analytics," *IEEE Access*, vol. 11, pp. 139097–139111, 2023.

[36] H. Pan, S. Guan, X. Zhao, and Y. Xue, "Edge computing-based vehicle detection in intelligent transportation systems," *Computing and Informatics*, vol. 42, no. 6, pp. 1339–1359, 2023.

[37] M. Ghasemi, Y. Fu, X. Ouyang, P. Wang, M. K. Turkcan, J. Tavori *et al.*, "Real-time video analytics for urban safety: Deployment over edge and end devices," in *Proceedings of the Tenth ACM/IEEE Symposium on Edge Computing*, Dec. 2025, pp. 1–17.

[38] H. Rabbouch, F. Saâdaoui, and R. Mraihi, "A vision-based statistical methodology for automatically modelling continuous urban traffic flows," *Advanced Engineering Informatics*, vol. 38, pp. 392–403, 2018.

[39] Z. Luo, P. M. Jodoin, S. Z. Su, S. Z. Li, and H. Larochelle, "Traffic analytics with low-frame-rate videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 878–891, 2016.

[40] V. C. Maha Vishnu, M. Rajalakshmi, and R. Nedunchezhian, "Intelligent traffic video surveillance and accident detection system with dynamic traffic signal control," *Cluster Computing*, vol. 21, no. 1, pp. 135–147, 2018.

[41] A. Nasry, A. Ezzahout, and F. Omary, "People tracking in video surveillance systems based on artificial intelligence," *Journal of Automation, Mobile Robotics and Intelligent Systems*, pp. 59–68, 2023.

[42] J. M. Alanazi, "Effective Machine-Learning Based Traffic Surveillance Moving Vehicle Detection," *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, vol. 14, no. 1, pp. 95–105, 2023.

[43] U. Bhanja, A. Mohanty, and S. Mahapatra, "Fuzzy logic-based accident detection system for vehicular ad hoc networks: a prototype implementation," *Wireless Personal Communications*, vol. 138, no. 1, pp. 291–320, 2024.

[44] K. SP, P. Mohandas, and S. CS, "Smart junction: advanced zone-based traffic control system with integrated anomaly detector," *Annals of operations research*, vol. 340, no. 1, pp. 479–506, 2024.

[45] I. Benallou, A. Azmani, and M. Azmani, "Evaluation of the accidents risk caused by truck drivers using a fuzzy Bayesian approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[46] D. Tian, C. Zhang, X. Duan, and X. Wang, "An automatic car accident detection method based on cooperative vehicle infrastructure systems," *IEEE Access*, vol. 7, pp. 127453–127463, 2019.

[47] L. Zhu, B. Wang, Y. Yan, S. Guo, and G. Tian, "A novel traffic accident detection method with comprehensive traffic flow features extraction," *Signal Image Video Process.*, 2022.

[48] O. I. Aboulola, "Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique," *PLoS one*, vol. 19, no. 4, p. e0300640, 2024.

[49] B. Kaur and J. Bhattacharya, "A convolutional feature map-based deep network targeted towards traffic detection and classification," *Expert Systems with Applications*, vol. 124, pp. 119–129, 2019.

[50] A. Pashaei, M. Ghatee, and H. Sajedi, "Convolution neural network joint with mixture of extreme learning machines for feature extraction and classification of accident images," *Journal of Real-Time Image Processing*, vol. 17, no. 4, pp. 1051–1066, 2020.

[51] K. Kumaran Santhosh, D. P. Dogra, P. P. Roy, and A. Mitra, "Vehicular Trajectory Classification and Traffic Anomaly Detection in Videos Using a Hybrid CNN-VAE Architecture," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11891–11902, Aug. 2022, doi: 10.1109/TITS.2021.3108504.

[52] Y. Xu *et al.*, "TAD: A Large-Scale Benchmark for Traffic Accidents Detection From Video Surveillance," *IEEE Access*, vol. 13, pp. 2018–2033, 2025, doi: 10.1109/ACCESS.2024.3522384.

[53] X. Wu and T. Li, "A deep learning-based car accident detection approach in video-based traffic surveillance," *Journal of Optics*, vol. 53, no. 4, pp. 3383–3391, 2024.

[54] A. Verma and M. Khari, "Vision-based accident anticipation and detection using deep learning," *IEEE Instrumentation & Measurement Magazine*, vol. 27, no. 3, pp. 22–29, 2024.

[55] S. Giriprasad, G. Ravikumar, S. Gokul, and S. D. Govardhan, "Highly efficient anomaly detection in traffic surveillance videos with optimized interference tolerant fast convergence zeroing neural network," *Signal Processing: Image Communication*, vol. 146, p. 117553, 2026, doi: 10.1016/j.image.2026.117553.

[56] S. Natha, F. A. Jokhio, M. Laghari, M. Siraj, S. A. Alsaif, U. Ashraf, and A. Ali, "A scalable and generalized deep ensemble model for road anomaly detection in surveillance videos," *Comput. Mater. Contin*, vol. 81, pp. 3707–3729, 2024.

[57] A. A. Pai, K. S. Harini, D. Giridhar, and S. Rangaswamy, "Real-Time Accident Detection from Closed Circuit Television and Suggestion of Nearest Medical Amenities," *IJITCS*, vol. 15, no. 6, pp. 15–23, 2023.

[58] Haryanto and O. Vaculín, "YoFlow Method for Scenario-Based Automatic Accident Detection," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 7, pp. 61–73, 2026, doi: 10.1109/OJITS.2025.3639557.

[59] S. Basheer, C. Biswas, D. Priyatham, S. A. Karim, and A. Rehan, "Enhancing Car Accident Detection: A Fuzzy Logic-Based Evaluation of Faster R-CNN and YOLOV8," in *Proceedings of 5th International Conference on Communication, Computing and Electronics Systems, ICCCES 2026*, 2026, pp. 515–522, doi: 10.1109/ICCCES62661.2026.11436489.

[60] D. T. Mane, S. Sangve, S. Kandhare, S. Mohole, S. Sonar, and S. Tupare, "Real-time vehicle accident recognition from traffic video surveillance using YOLOV8 and OpenCV," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5s, pp. 250–258, 2023.

[61] Y. Zhang and Y. Sung, "Traffic accident detection method using trajectory tracking and influence maps," *Mathematics*, 2023.

[62] M. I. Basheer Ahmed, R. Zaghdoud, M. S. Ahmed, R. Sendi, S. Alsharif, J. Alabdulkarim *et al.*, "A real-time computer vision based approach to detection and classification of traffic incidents," *Big data and cognitive computing*, vol. 7, no. 1, p. 22, 2023.

[63] T. K. Vijay, D. P. Dogra, H. Choi, G. Nam, and I. J. Kim, "Detection of road accidents using synthetically generated multi-perspective accident videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 1926–1935, 2022.

[64] N. Wang, Q. Deng, Z. Jiao, and Z. Zhong, "A fast traffic accident recognition method based on edge computing and deep neural network," in *Proceedings of the Romanian Academy Series A-Mathematics Physics Technical Sciences Information Science*, vol. 23, no. 1, Jan. 2022, pp. 69–78.

[65] S. Robles-Serrano, G. Sanchez-Torres, and J. Branch-Bedoya, "Automatic detection of traffic accidents from video using deep learning techniques," *Computers*, vol. 10, no. 11, p. 148, 2021.

[66] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video," *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, vol. 6, no. 2, pp. 1–28, 2020.

[67] Z. Lu, W. Zhou, S. Zhang, and C. Wang, "A new video-based crash detection method: balancing speed and accuracy using a feature fusion deep learning framework," *Journal of advanced transportation*, vol. 2020, no. 1, p. 8848874, 2020.

[68] P. Kalpana, G. Sowmiya, C. S. Sri, and S. Sivapriya, "Road traffic accident detection based on YOLOv8 and Byte Track," in *AIP Conference Proceedings*, vol. 3204, no. 1, Feb. 2025, p. 040013.

[69] M. Karthikeyan, P. R. Kshirsagar, T. K. Tak, K. Lalit, K. Upreti, and R. Jain, "AADS: An Automated Accident Detection and Nighttime surveillance system using fine-tuned YOLOV10 deep learning techniques," in *2025 International Conference on Intelligent Control, Computing and Communications (IC3)*, Feb. 2025, pp. 1351–1356.

[70] G. P. D. Estrada, A. J. T. Go, J. R. E. Martinez, J. A. M. T. Natividad, and J. P. Cempron, "Temporal Deep Learning Anomaly Detection for Pedestrian Accident Detection in Video Surveillance," in *2025 31st International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, Aug. 2025, pp. 238–242.

[71] J. M. Quek, H. L. Khor, and J. Ooi, "YOLOv5 for Accident Detection: Review and a Case for MSRCR-Based Image Enhancement," in *2025 IEEE 17th Malaysia International Conference on Communication (MICC)*, Aug. 2025, pp. 1–6.

[72] A. Padma, Y. S. A. Kurivella, V. Duggineni, and N. Kothamasu, "AI-Based Traffic Detection: Enhancing Road Safety and Efficiency Using Deep Learning," in *2025 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, Aug. 2025, pp. 1–8.

[73] V. K. G. Kalaiselvi and R. Ranjana, "Enhanced road safety through intelligent accident detection and automated emergency intimations," in *2025 International Conference on Computing and Communication Technologies (ICCCT)*, Apr. 2025, pp. 1–6.

[74] S. Yang, M. Abdel-Aty, Z. Islam, and D. Wang, "Real-time crash prediction on express managed lanes of Interstate highway with anomaly detection learning," *Accident Analysis & Prevention*, vol. 201, p. 107568, 2024.

[75] V. A. Kotkar and V. Sucharita, "Scalable anomaly detection framework in video surveillance using keyframe extraction and machine learning algorithms," *Journal of*

Advanced Research in Dynamical and Control Systems, vol. 12, no. 7, pp. 395–408, 2020.

[76] P. Kadam, D. Vora, S. Patil, S. Mishra, and V. Khairnar, "Behavioral profiling for adaptive video summarization: From generalization to personalization," *MethodsX*, vol. 13, p. 102780, 2024, doi: 10.1016/j.mex.2024.102780.

[78] J. Xu, Z. Sun, and C. Ma, "Crowd aware summarization of surveillance videos by deep reinforcement learning," *Multimed Tools Appl*, vol. 80, pp. 6121–6141, 2021, doi: 10.1007/s11042-020-09888-1.

[79] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3d spatio-temporal u-net," *IEEE Trans Image Process*, vol. 31, pp. 1573–1586, 2022.

[80] G. Wang, X. Wu, and J. Yan, "Progressive reinforcement learning for video summarization," *Information Sciences*, vol. 655, p. 119888, 2024, doi: 10.1016/j.ins.2023.119888.

[81] I. Taheri Sarteshnizi, M. Sarvi, S. A. Bagloee, and N. Nassir, "Temporal pattern mining of urban traffic volume data: a pairwise hybrid clustering method," *Transport metrica B: Transport Dynamics*, vol. 11, no. 1, p. 2185496, 2023.

[82] H. Rabbouch, F. Saâdaoui, and R. Mraihi, "Unsupervised video summarization using cluster analysis for automatic vehicles counting and recognizing," *Neurocomputing*, vol. 260, pp. 157–173, 2017, doi: 10.1016/j.neucom.2017.04.026.

[83] Q. Qi, X. Wang, T. Hou, Y. Yan, and H. Wang, "FastVOD-Net: A real-time and high-accuracy video object detector," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20926–20942, 2022.

[84] J. Chen, J. Pu, B. Yin, R. Zhang, and J. J. Wu, "TA-NET: Empowering Highly Efficient Traffic Anomaly Detection Through Multi-Head Local Self-Attention and Adaptive Hierarchical Feature Reconstruction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 12372–12384, 2024.

[85] H. Li, Z. Li, X. Wang, Z. Pan, and Z. Qu, "A Real-Time Automatic Traffic Event Detection Method Based on Temporal Convolutional Autoencoder Network," *China*

Journal of Highway and Transport, vol. 35, no. 6, pp. 265–276, 2022, doi: 10.19721/j.cnki.1001-7372.2022.06.022.

[86] S. Sneha, S. Sriranjini, and M. Balaji, "An IoT-Enabled Intelligent Traffic Management System with CNN-Based Accident Detection and Dynamic Rerouting Using K-Means and Dijkstra's Algorithm," in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, Jan. 2025, pp. 423–430.

[87] M. Zohaib, M. Asim, and M. ELAffendi, "Enhancing emergency vehicle detection: A deep learning approach with multimodal fusion," *Mathematics*, vol. 12, no. 10, p. 1514, 2024.

[88] N. Vasundhara, L. Pallavi, K. R. Reddy, A. Rajesh, and A. M. Rajeswari, "Integrating machine learning algorithms to analyze and determine traffic projections for intelligent transportation systems," in *AIP Conference Proceedings*, vol. 3342, no. 1, Sep. 2025, p. 030047.

[89] J. Sravanthi, V. Bhukya, R. Yeligeti, S. Elugubanti, M. A. Mohiuddin, M. M. Adnan *et al.*, "Enhanced traffic violation monitoring system utilizing deep neural networks for accurate offense identification and license plate recognition," in *AIP Conference Proceedings*, vol. 3263, no. 1, Aug. 2025, p. 080009.

[90] M. Q. Kheder and A. A. Mohammed, "Transfer learning based traffic light detection and recognition using CNN inception-V3 model," *Iraqi Journal of Science*, pp. 6258–6275, 2023.

[91] A. Khan, S. Khan, B. Hassan, and Z. Zheng, "CNN-based smoker classification and detection in smart city application," *Sensors*, vol. 22, no. 3, p. 892, 2022.

[92] B. R. Pilli, S. Nagarajan, and P. Devabalan, "Detecting the vehicle's number plate in the video using deep learning performance," *Rigeo*, vol. 11, no. 5, 2021.

[93] J. Hu, S. Abubakar, S. Liu, X. Dai, G. Yang, and H. Sha, "Near-infrared road-marking detection based on a modified faster regional convolutional neural network," *Journal of Sensors*, vol. 2019, no. 1, p. 7174602, 2019.

[94] M. N. Ubaidillah, A. Nugraha, A. Luthfiarta, E. Y. Hidayat, C. L. Buana, and R. Marendratama, "Performance Comparison of YOLOv8 and YOLOv11 Enhanced by Grey Wolf Optimizer for Motorcycle Helmet Detection in Traffic Surveillance," in

2025 International Seminar on Application for Technology of Information and Communication (iSemantic), Sep. 2025, pp. 351–356.

[95] M. Sathyamoorthy, V. Rajasekar, S. Krishnamoorthi, and D. Pamucar, "Ensemble deep learning approach for traffic video analytics in edge computing," *Scientific Reports*, 2026.

[96] H. Kim, A. Khudoyberdiev, S. SR Garnaik, A. Bhattacharya, and J. Ryoo, "CLOUD-CODEC: A New Way of Storing Traffic Camera Footage at Scale," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 8, pp. 1–28, 2025.

[97] Y. Yang, X. Chen, J. Wang, Y. Dong, K. Qie, and Z. Yuan, "Enhancing vision-based traffic crash detection performance consistency across day-night scenes: A depth-aware and domain-adaptive network," *Accident Analysis & Prevention*, vol. 228, p. 108405, 2026.

[98] S. Khanam, M. Sharif, M. Raza, W. Ishaq, M. Fayyaz, and S. Kadry, "Anomaly recognition in surveillance based on feature optimizer using deep learning," *PLoS One*, vol. 20, no. 5, p. e0313692, 2025.

[99] A. Shaer, A. Fielbaum, and D. Levinson, "Detecting dangerous driving via computer vision: Linking video-based indicators to road crashes," *Journal of Transportation Safety & Security*, pp. 1–20, 2025.

[100] A. Vrochidis, S. Panagiotidis, S. Parcharidis, N. Dimitriou, P. Papaioannou, E. Milolidaki *et al.*, "Leveraging 5G networks for event video summarization via multimodal analysis," *Discover Artificial Intelligence*, 2025.

[101] S. B. Veesam and A. R. Satish, "An Empirical Taxonomy of Video Summarization Model From a Statistical Perspective," *IEEE Access*, vol. 12, pp. 173850–173866, 2024.

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

[103] S. Natha, F. Ahmed, M. Siraj, M. Lagari, M. Altamimi, and A. A. Chandio, "Deep BiLSTM attention model for spatial and temporal anomaly detection in video surveillance," *Sensors*, vol. 25, no. 1, p. 251, 2025.

[104] J. Lin, S.-H. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Computers & Electrical Engineering*, vol. 97, p. 107618, 2022, doi: 10.1016/j.compeleceng.2021.107618.

[105] H. Liu, X. Hu, G. Sun, W. Zhang, J. Zhan, H. Fang *et al.*, "Intelligent traffic accident detection system in complex dynamic scenarios based on the dual-stream spatiotemporal-fusion model," *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112675, 2025.

[106] W. Zhou, L. Yang, L. Zhao, R. Zhang, Y. Cui, H. Huang *et al.*, "Vision technologies with applications in traffic surveillance systems: A holistic survey," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–47, 2025.

[107] K. Zhang, W. L. Chao, F. Sha, and K. Grauman, "Video Summarization with Long Short-Term Memory," in *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*, vol. 9911, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, Cham, 2016, doi: 10.1007/978-3-319-46478-7_47.

[108] J. Zhong, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709–1717, 2019.

[109] J. An and S. Cho, "Variational Autoencoder based Anomaly Detection using Reconstruction Probability," 2015.

[110] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly-supervised video summarization using variational encoder-decoder and web prior," in *ECCV*, 2018, pp. 184–200.

[111] A. Almutairi, A. F. Al Asmari, F. Alanazi, T. Alqubaysi, and A. Armghan, "Deep learning based predictive models for real time accident prevention in autonomous vehicle networks," *Scientific Reports*, vol. 15, no. 1, p. 20844, 2025.

[112] A. Zahid, T. Qasim, N. Bhatti, and M. Zia, "A data-driven approach for road accident detection in surveillance videos," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17217–17231, 2024.

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

[114] T. Iqbal and S. Qureshi, "Reconstruction probability-based anomaly detection using variational auto-encoders," *International Journal of Computers and Applications*, vol. 45, no. 3, pp. 231–237, 2022, doi: 10.1080/1206212X.2022.2143026.

[115] W. Li, L. Huang, and X. Lai, "A deep learning framework for traffic accident detection based on improved YOLO11," *Vehicles*, vol. 7, no. 3, p. 81, 2025.

[116] W. Yu and Q. 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*, vol. 115, p. 103115, 2022, doi: 10.1016/j.jag.2022.103115.

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

[118] B. Wang and C. Yang, "Video Anomaly Detection Based on Convolutional Recurrent AutoEncoder," *Sensors*, vol. 22, no. 12, p. 4647, 2022, doi: 10.3390/s22124647.

[119] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 879–887, 2019, doi: 10.1109/TITS.2018.2835308.

[120] J. Ashraf, A. D. Bakhshi, N. Moustafa, H. Khurshid, A. Javed, and A. Beheshti, "Novel Deep Learning-Enabled LSTM Autoencoder Architecture for Discovering Anomalous Events From Intelligent Transportation Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4507–4518, 2021, doi: 10.1109/TITS.2020.3017882.

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

[122] S. Hameed, J. Amin, M. A. Anjum, and M. Sharif, "Suspicious activities detection using spatial-temporal features based on vision transformer and recurrent neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 9, pp. 3379–3391, 2024.

[123] X. Chen, H. Xu, M. Ruan, M. Bian, Q. Chen, and Y. Huang, "SO-TAD: A surveillance-oriented benchmark for traffic accident detection," *Neurocomputing*, vol. 618, p. 129061, 2025.

[124] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised Object-Level Video Summarization with Online Motion Auto-Encoder," *Pattern Recognition Letters*, vol. 130, pp. 376–385, 2020, doi: 10.1016/j.patrec.2018.07.030.

[125] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

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

[127] S.-H. Zhong, J. Lin, J. Lu, A. Fares, and T. Ren, "Deep Semantic and Attentive Network for Unsupervised Video Summarization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 2, Article 55, May 2022, doi: 10.1145/3477538.

[128] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6479–6488.

[129] A. Srinivasan, A. Srikanth, H. Indrajit, and V. Narasimhan, "A Novel Approach for Road Accident Detection using DETR Algorithm," in *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2020, pp. 75–80, doi: 10.1109/IDSTA50958.2020.9263703.

[130] C. Wang, Y. Dai, W. Zhou, and Y. Geng, "A Vision-Based Video Crash Detection Framework for Mixed Traffic Flow Environment Considering Low-Visibility Condition," *Journal of Advanced Transportation*, 2020, doi: 10.1155/2020/9194028.

[131] S. W. Khan, Q. Hafeez, M. I. Khalid, R. Alroobaea, S. Hussain, J. Iqbal, J. Almotiri, and S. S. Ullah, "Anomaly Detection in Traffic Surveillance Videos Using Deep Learning," *Sensors*, vol. 22, no. 17, p. 6563, 2022, doi: 10.3390/s22176563.

[132] K. Pawar and V. Attar, "Deep learning based detection and localization of road accidents from traffic surveillance videos," *ICT Express*, 2021, doi: 10.1016/j.icte.2021.11.004.

[133] A. R. Pathak and A. C. Elster, "Applying Transfer Learning to Traffic Surveillance Videos for Accident Detection," in *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*, Halden, Norway, 2022, pp. 1–7, doi: 10.1109/ICAPAI55158.2022.9801568.

[134] 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," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20601–20613, Nov. 2022, doi: 10.1109/TITS.2022.3170648.

[135] V. A. Adewopo and N. Elsayed, "Smart City Transportation: Deep Learning Ensemble Approach for Traffic Accident Detection," *IEEE Access*, vol. 12, pp. 59134–59147, 2024, doi: 10.1109/ACCESS.2024.3387972.

[136] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Event Detection on Roads Using Perceptual Video Summarization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 2944–2954, Sept. 2018, doi: 10.1109/TITS.2017.2769719.

[137] T. Kosambia and J. Gheewala, "Video Synopsis for Accident Detection using Deep Learning Technique," in *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, May 2021, doi: 10.2139/ssrn.3851250.

[138] 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," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24116–24125, Dec. 2022, doi: 10.1109/TITS.2022.3198595.

[139] M. Tahir, Y. Qiao, N. Kanwal, B. Lee, and M. N. Asghar, "Privacy Preserved Video Summarization of Road Traffic Events for IoT Smart Cities," *Cryptography*, vol. 7, no. 1, p. 7, 2023, doi: 10.3390/cryptography7010007.

[140] N. Saxena and M. N. Asghar, "YOLOv5 for Road Events Based Video Summarization," in *Intelligent Computing. SAI 2023. Lecture Notes in Networks and Systems*, vol. 739, K. Arai, Ed. Springer, Cham, 2023, doi: 10.1007/978-3-031-37963-5_69.

List of 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. 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.