Received: July 29, 2025. Revised: November 11, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Explainable AI for Extraction and Analysis of Spatial and Temporal Features in Traffic Surveillance Videos

Vasanth Kumar N T[1]* **Geetha Kiran A[1]**

[1]*Department of Computer Science and Engineering, Malnad College of Engineering, Hassan Affiliated to Visvesvaraya Technological University, Belagavi, India*
* Corresponding author's Email: agk@mcehassan.ac.in

Abstract: This paper introduces EXTRA (EXplainable Traffic Analysis), a novel explainable AI methodology for real-time accident prediction in traffic surveillance videos. EXTRA integrates advanced spatial and temporal feature extraction with context-aware decision making, combining CNNs (Convolutional Neural Networks), LSTMs (Long Short-Term Memory), and GNN (Graph Neural Networks) . Our approach uniquely focuses on generating human-interpretable explanations for its predictions. Evaluated on the CADP dataset and validated through cross-dataset evaluation on TADS (2024) and Kaggle CCTV datasets, EXTRA demonstrates state-of-the-art performance with an Average Precision of 87.3% on CADP and maintains robust generalization with 76.8% accuracy on TADS and 72.1% on Kaggle CCTV datasets. The algorithm achieves real-time processing at 28.3 FPS for 1280x720 resolution inputs while preserving explainability across diverse surveillance environments. EXTRA provides high-fidelity explanations, with an Explanation Fidelity of 89.7% on CADP and consistent interpretability scores above 3.5 across all tested datasets.

Keywords: Traffic surveillance, Spatial features, Temporal features, CADP dataset, Explainable AI, Image processing, LSTM networks, Interpretability.

1. Introduction

Traffic accidents remain a significant global concern, posing substantial risks to human life and property. Despite ongoing improvements in road infrastructure and vehicle safety systems, the complex and dynamic nature of traffic environments continues to challenge our ability to predict and prevent accidents effectively. In recent years, the advent of advanced machine learning techniques, particularly deep learning, has opened new avenues for enhancing traffic safety through more accurate and timely accident prediction. The field of traffic accident prediction has seen a shift from traditional statistical methods to more sophisticated approaches that can handle the complexity and volume of modern traffic data. Deep learning models have demonstrated superior performance in capturing intricate patterns and relationships within traffic data, leading to improved prediction accuracy. The objective of this research is to address this challenge by developing an explainable AI framework for traffic accident prediction, with a focus on extracting and analyzing spatial and temporal features from traffic surveillance videos. By leveraging the CADP (CCTV Accident Dataset Public) [7], a comprehensive dataset of traffic camera footage, we aim to create a model that not only predicts accidents with high accuracy but also provides clear, interpretable explanations for its predictions.

A. Relevance of the problem

The relevance of traffic accident prediction and prevention in urban planning and management cannot be overstated. Traffic accidents continue to be a major cause of fatalities and injuries worldwide, with significant economic and social impacts. In 2016 alone, there were 40,200 deaths due to car and motor vehicle accidents on national roads in the United States [1]. This staggering figure underscores

the urgent need for more effective traffic safety measures and accident prevention strategies. Traffic surveillance plays a crucial role in addressing this challenge. As urban populations grow and traffic densities increase, the complexity of managing traffic flow and ensuring road safety becomes more pronounced. Traffic cameras, which provide broad, fixed, and objective views of multiple vehicles and their environment [2], have become an invaluable tool in this context. They offer a rich source of real-time data that can be leveraged for proactive traffic management and accident prevention.

However, the utilization of this data presents significant challenges:

1) Volume and Complexity: Traffic surveillance cameras generate vast amounts of video data. Processing and analyzing this data in real-time to extract meaningful insights is a formidable task [3].

2) Feature Extraction: Identifying and extracting relevant features from video data that can accurately predict accident risk is complex. These features may include spatial elements (such as road layout and vehicle positions) and temporal aspects (like traffic flow patterns and vehicle trajectories) [4].

3) Interpretability: While advanced machine learning models, particularly deep learning approaches, have shown promise in processing this complex data [5], they often lack interpretability. This "black box" nature can be problematic when decisions based on these models have significant safety implications [6].

4) Real-time Processing: The need for real-time analysis to enable immediate interventions adds another layer of complexity to the problem [7].

5) Rare Event Prediction: Traffic accidents, while devastating, are relatively rare events. This imbalance in data makes accurate prediction particularly challenging [8].

Moreover, as cities move towards smart transportation systems and autonomous vehicles become more prevalent, the need for explainable and reliable traffic analysis tools becomes even more critical. These systems will rely heavily on accurate, real-time understanding of traffic conditions and risk factors [9].

Figure. 1 Sample frames from CADP dataset

In light of these considerations, research into explainable AI methods for traffic accident prediction using surveillance data is not just relevant, but essential for the future of urban traffic management and road safety.

B. Statement of the problem

The central challenge this research addresses is the lack of efficient and explainable methods for extracting and analyzing spatial and temporal features in traffic surveillance videos for accident prediction. Current approaches face several key issues:

1) The need for models that can generalize across diverse traffic environments [2].

2) Difficulty in extracting interpretable spatial features from complex traffic scenes [4].

3) The "black box" nature of deep learning models, which hinders their adoption in safety-critical applications [5].

4) Challenges in capturing and explaining temporal dynamics of traffic patterns [6].

5) The need for integrating spatial and temporal analyses in a unified, explainable framework [7].

6) Requirements for real-time processing and analysis to enable timely interventions [8].

7) Imbalanced datasets, as accidents are rare events compared to normal traffic flow [9].

This research aims to develop an explainable AI framework that efficiently extracts and analyses both spatial and temporal features from traffic surveillance videos. The goal is to predict accidents accurately while providing interpretable insights into contributing factors, thereby enhancing the trustworthiness and adoptability of AI-driven traffic management systems.

C. Review - State of the art

Real-time crash risk prediction has gained significant attention in recent years, with a focus on enhancing road safety through advanced traffic management systems. This review focuses on the state-of-the-art approaches, particularly emphasizing deep learning methods for real-time crash risk prediction on urban arterials.

1) Traditional Approaches: Early studies on real-time crash risk prediction primarily focused on freeways rather than urban arterials due to the more complex traffic environment of the latter [1]. Traditional statistical methods, such as logistic regression and log-linear models, were commonly used [2]. For instance, Abdel-Aty et al. (2004)

employed matched case-control logistic regression for freeway crash prediction [3].

2) Machine Learning Advancements: As data availability improved, machine learning methods gained popularity. Theofilatos et al. (2019) compared various machine learning and deep learning approaches for real-time crash prediction, finding that deep learning methods, particularly Deep Neural Networks (DNN), outperformed traditional techniques [4]. Other machine learning methods like Support Vector Machine (SVM) and Random Forest have also been applied successfully [5, 6].

3) Deep Learning Innovations: Recent years have seen a surge in deep learning applications for traffic safety analysis:

1) Recurrent Neural Networks (RNN): Long Short-Term Memory (LSTM) networks have shown remarkable effectiveness in capturing temporal dynamics of traffic flow. Yuan et al. (2019) utilized LSTM for real-time crash risk prediction, achieving significantly higher sensitivity compared to traditional models [7].

2) Convolutional Neural Networks (CNN): CNNs have been effective in extracting spatial features from traffic data. Ma et al. (2017) used CNN for traffic speed prediction, demonstrating its ability to capture spatial and temporal dependencies [8].

3) Hybrid Models: Combining different neural network architectures has shown promising results. The LSTMCNN model, which benefits from both LSTM's ability to capture long-term dependencies and CNN's capability to extract time-invariant features, has been particularly effective [9].

4) Data Integration and Preprocessing: A key advancement has been the integration of diverse data sources. Recent studies have combined traffic flow data, signal timing information, weather conditions, and historical accident reports [10]. This multi-modal approach allows for a more comprehensive understanding of accident risk factors. To address the imbalanced nature of crash data (where noncrash events significantly outnumber crash events), techniques like Synthetic Minority Over-sampling Technique (SMOTE) have been employed [11].

5) Real-time Processing and Explainability: The implementation of real-time processing in deep learning models has been crucial for proactive road safety measures. Recent research has focused on developing explainable AI techniques specifically for traffic accident prediction, aiming to provide transparent reasoning behind predictions [12].

D. Challenges and Research Opportunities

Despite significant advancements, several challenges remain:

1) Handling the rarity of accident events in large datasets.

2) Ensuring real-time performance for immediate safety interventions.

3) Balancing model complexity with interpretability for practical application in traffic management systems.

4) Developing models that can generalize across different urban environments and traffic conditions.

Research opportunities include the integration of edge computing for more efficient real-time processing, development of more sophisticated spatio-temporal models, and exploration of transfer learning approaches to enhance model generalization across different traffic environments.

E. Contributions

This research presents EXTRA, a novel explainable AI framework for real-time crash risk prediction on urban arterials. Our key contributions include:

1) Development of a framework that integrates spatial and temporal feature analysis for accident prediction

2) Enhancement of object detection and tracking capabilities through advanced computer vision techniques

3) Implementation of comprehensive explainability metrics.

4) Achievement of real-time processing capabilities while maintaining a practical model size.

5) Thorough ablation studies demonstrating the effectiveness of each component in the EXTRA framework.

6) Comprehensive comparative analysis with state-of-the-art methods.

These contributions represent significant advancements in real-time crash risk prediction, offering a robust, accurate, and explainable method for enhancing road safety on urban arterials.

This study contributes to the field by:

- Developing novel techniques for extracting explainable spatial features from traffic video data.
- Investigating methods for explainable temporal feature extraction to capture the dynamic nature of traffic flows.
- Integrating spatial and temporal analysis within an XAI framework to provide a comprehensive understanding of accident risk factors.

By combining the predictive power of deep learning with the transparency of explainable AI, our research aims to advance the state of the art in traffic

accident prediction and contribute to the development of more effective, trustworthy traffic safety systems.

The paper is organized as follows. Section 2 discusses model building, followed by Section 3 which details the proposed EXTRA algorithm. Section 4 presents experimental results and comparative analysis. Finally, Section 5 concludes the paper and suggests future research directions.

2. Model building with XAI

This section presents a comprehensive framework for explainable AI in traffic surveillance video analysis, focusing on accident prediction and interpretation. Our model integrates advanced spatial and temporal feature extraction techniques with context-aware decision making and explainable outputs.

A. Model Overview

The proposed model consists of five main components:

1) Spatial Feature Extraction
2) Temporal Feature Analysis
3) Context Mining
4) Multi-Object Interaction Modeling
5) Explainable Decision Making and Output Generation

Let V = {f1, f2, . . . , fn} be a video sequence of n frames, where each frame ft ∈ RH×W×3 represents an RGB image of height H and width W.

B. Detailed Model Architecture

EXTRA's architecture integrates multiple specialized components to achieve robust crash risk prediction while maintaining explainability. The framework processes input video frames through five key components: spatial feature extraction using convolutional neural networks, temporal feature analysis via LSTM networks, context mining for environmental understanding, multi-object interaction modeling through graph neural networks, and an explainable decision-making module.

The architecture processes each video frame f_t sequentially, first extracting spatial features to capture the visual characteristics of the scene. These features then feed into both a temporal analysis module that tracks evolution over time and a context mining module that extracts environmental factors.

Simultaneously, a graph neural network models the interactions between detected objects.

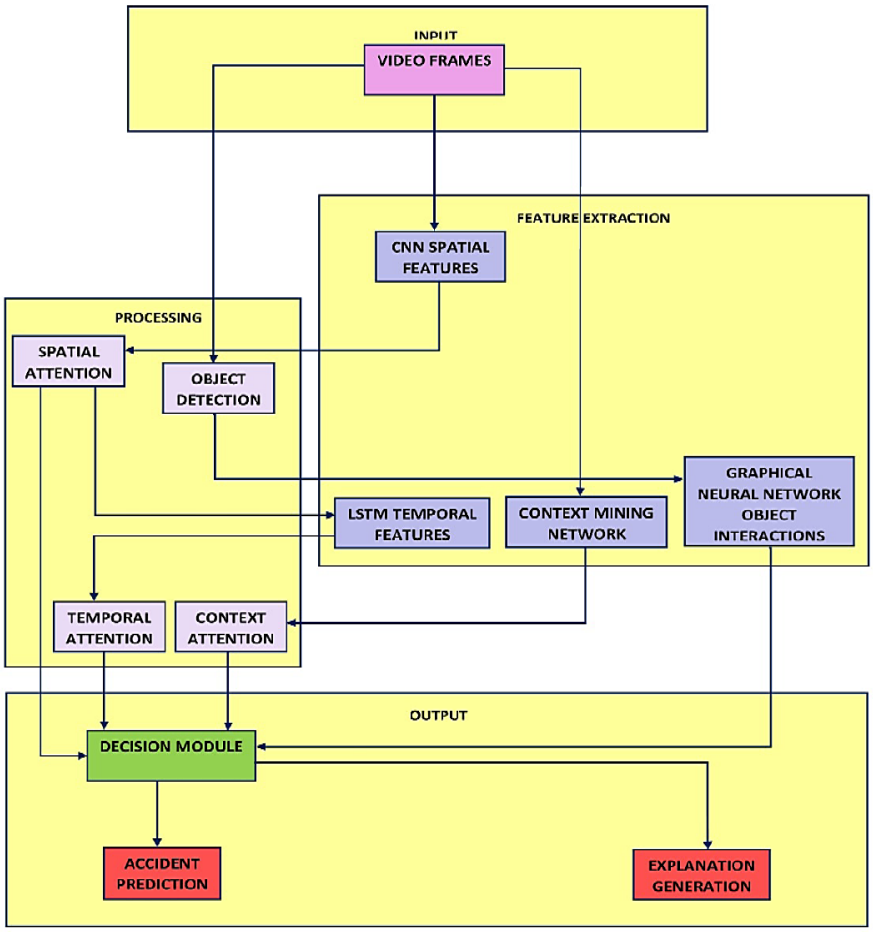


Figure. 2 Detailed Model Architecture Diagram

Finally, all these features converge in an explainable decision-making module that outputs both the crash risk prediction and corresponding explanations for its decisions.

This modular design allows each component to focus on specific aspects of the crash prediction task while working together to provide comprehensive analysis. We detail each component in the following subsections.

C. Spatial Feature Extraction

We employ a deep convolutional neural network φ to extract spatial features:

$$St = \varphi(ft) \tag{1}$$

where $S_t \in R^{d\times m\times m}$ represents the spatial features, d is the feature dimension, and m × m is the spatial resolution.

The CNN architecture is defined as a series of convolutional layers:

$$\varphi(ft) = \sigma_L(W_L * \sigma_{L-1}(W_{L-1} * \dots \sigma_1(W_1 * f_t + b_1) \cdots + b_{L-1}) + b_L) \tag{2}$$

where W_L and b_L are the weights and biases of the L-th layer, σ_L is the activation function, and $*$ denotes the convolution operation.

D. Temporal Feature Analysis

To capture temporal dynamics, we utilize a recurrent neural network ψ, specifically a Long Short-Term Memory (LSTM) network:

$$H_t = \psi(S_t, H_{t-1}) \tag{3}$$

where $H_t \in R^h$ is the hidden state at time t, and h is the hidden state dimension.

The LSTM update equations are:

$$i_t = \sigma(W_i[H_{t-1}, S_t] + b_i) \tag{4}$$

$$f_t = \sigma(W_f[H_{t-1}, S_t] + b_f) \tag{5}$$

$$o_t = \sigma(W_o[H_{t-1}, S_t] + b_o) \tag{6}$$

$$\tilde{c}_t = \tanh(W_c[H_{t-1}, S_t] + b_c) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{8}$$

$$H_t = o_t \odot \tanh(c_t) \tag{9}$$

where i_t, f_t, and o_t are the input, forget, and output gates respectively, c_t is the cell state, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

E. Context Mining

Context information C_t is extracted using a separate network θ:

$$C_t = \theta(f_t) \tag{10}$$

where $C_t \in Rc$ represents c-dimensional context features.

The context mining network θ is defined as:

$$\theta(f_t) = FC_K(\sigma_K(\dots FC_1(\sigma_1(CNN(f_t)))\dots)) \tag{11}$$

where CNN is a lightweight convolutional network, FC_k are fully connected layers, and σ_k are activation functions.

F. Multi-Object Interaction Modeling

To model interactions between multiple objects in the scene, we introduce a graph neural network (GNN) ξ:

$$G_t = \xi(B_t, S_t) \tag{12}$$

where $B_t = \{b_1, b_2, \dots, b_K\}$ is the set of K detected object bounding boxes, and $G_t \in R^{K\times g}$ represents the interaction features for each object.

The GNN update equation is:

$$g_i^{(l+1)} = \Gamma(g_i^l, \sum_{j\in N(i)} \Phi(g_i^l, g_j^l, e_{ij})) \tag{13}$$

where g_i^l is the feature of node *i* at layer *l*, $N(i)$ is the set of neighboring nodes, e_{ij} is the edge feature between nodes *i* and *j*, Φ is a message function, and Γ is an update function.

G. Explainable Decision Making and Output Generation

The final decision function δ combines all extracted features:

$$P(A_t), E_t = \delta(H_t, C_t, G_t) \tag{14}$$

where $P(A_t)$ is the probability of an accident at time *t*, and E_t is the corresponding explanation vector.

The decision function is formulated as:

$$P(A_t) = \sigma(W_p[H_t, C_t, \sum g_i] + b_p) \tag{15}$$

$$E_t = \text{softmax}(W_e[H_t, C_t, G_t] + b_e) \tag{16}$$

where W_p, W_e, b_p, b_e are learnable parameters, and E_t represents the importance of different factors in the decision- making process.

H. Model Training and Optimization

The model is trained end-to-end using a multi-task loss function:

$$L = \lambda_1 L_{\text{det}} + \lambda_2 L_{\text{pred}} + \lambda_3 L_{\text{exp}} \quad (17)$$

where L_{det} is the object detection loss, L_{pred} is the accident prediction loss, and L_{exp} is the explanation generation loss. The λ_i are weighting factors.

The optimization problem is formulated as:

$$\min \mathrm{E}_V[L(V;\Theta)] \quad (18)$$

where Θ represents all learnable parameters of the model, and the expectation is taken over all training videos V.

I. Theoretical Justification

The effectiveness of our model is based on the following theoretical considerations:

1). **Spatial Feature Extraction**: The CNN φ learns hi- erarchical features, capturing both low-level textures and high-level semantic information crucial for object detection and scene understanding.

2). **Temporal Feature Analysis**: The LSTM ψ models long-term dependencies in the video sequence, address- ing the challenge of predicting rare events like accidents.

3). **Context Mining**: Explicit modeling of context through θ enhances the model's ability to adapt to various traffic scenarios and environmental conditions.

4). **Multi-Object Interaction Modeling**: The GNN ξ cap-tures complex relationships between objects, crucial for understanding traffic dynamics and potential collision risks.

5). **Explainable Decision Making**: The decision function δ is designed to be interpretable, allowing for the generation of human-understandable explanations while maintaining high predictive accuracy.

3. Algorithm

This section presents a comprehensive description and analysis of our proposed algorithm for explainable AI in traffic surveillance video analysis, focusing on the extraction and interpretation of spatial and temporal features for accident prediction and explanation.

We introduce EXTRA (EXplainable TRaffic Analysis), a novel algorithm designed to process traffic surveillance video streams, predict potential accidents, and generate human- interpretable explanations. EXTRA integrates advanced computer vision techniques such as Explanable AI, deep learning models, and explainable AI principles to address the challenges of real-time traffic analysis.

A. Theoretical Foundations

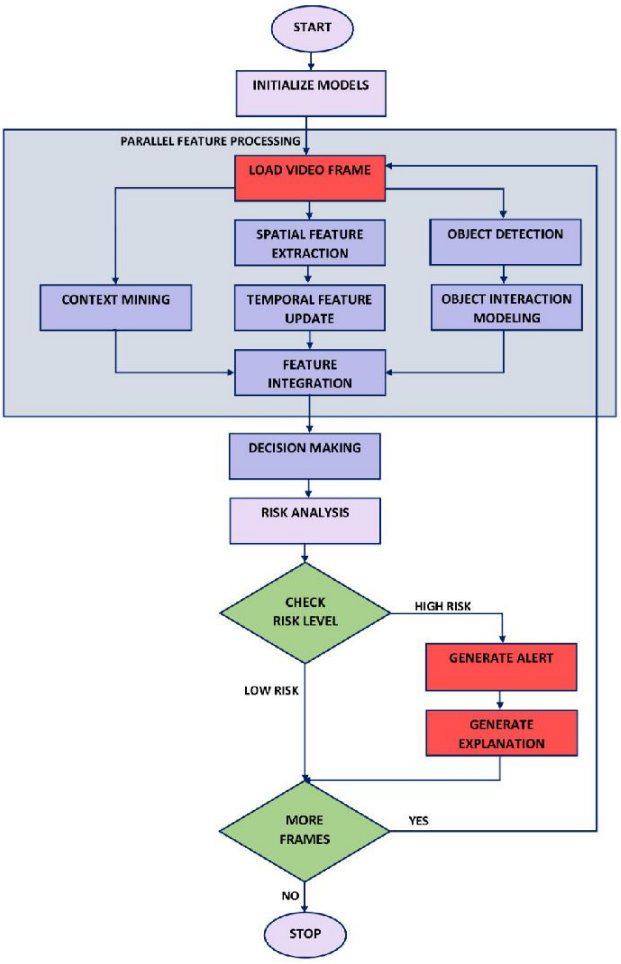


Figure. 3 EXplainable TRaffic Analysis

Before delving into the algorithmic details, we present the theoretical foundations underpinning EXTRA:

1) Spatio-Temporal Feature Learning: EXTRA builds upon the concept of spatio-temporal feature learning, which com- bines:

1. Convolutional Neural Networks (CNNs) for spatial feature extraction
2. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for temporal dependency modeling

The integration of CNNs and LSTMs is formulated as:

$$H_t = \mathrm{LSTM}(\mathrm{CNN}(f_t), H_{t-1}) \quad (19)$$

Where f_t is the input frame at time t, and H_t is the hidden state capturing both spatial and temporal information.

2) Graph Neural Networks for Object Interaction: We model object interactions using Graph Neural Networks (GNNs). Given a set of objects $O = \{o_1, \ldots, o_N\}$, we construct a graph $G =$

(V, E), where V represents objects and E represents their relationships.

The GNN update rule is:

$$h_i^{(k+1)} = \Gamma(h_i^k, X_j \in N(i)\ \Phi(h_i^k, h_j^k, e_{ij})) \quad (20)$$

where h^k is the feature of node i at layer k, $N(i)$ is the neighborhood of i, e_{ij} is the edge feature, Φ is a message function, and Γ is an update function.

3) Explainable AI Principles: We incorporate explainable AI principles through:

1. Attention mechanisms to highlight important spatial regions and temporal instances.
2. Feature importance scoring to identify critical factors in decision-making.
3. Rule extraction to generate human-readable explanations.

The Explanation Fidelity (EF) metric is formally defined as:

EF =(1/N) × Σ(i=1 to N) [correlation(A_model(x_i), A_explanation(x_i))]

where:

- A_model(x_i) = attention weights from the model's decision process
- A_explanation(x_i) = attention weights from the explanation generation
- N = number of test samples
- correlation() = Pearson correlation coefficient

B. Detailed Algorithm Description

Algorithm 1 EXTRA (EXplainable TRaffic Analysis)

Input: *Video sequence* $V = \{f_1, f_2, \ldots, f_n\}$
Output: *Accident predictions* $\{P(A_t)\}$ *and explanations* $\{E_t\}$

Initialize models:
- Spatial feature extractor φ
- Temporal feature processor ψ
- Context mining network θ
- Object interaction modeler ξ
- Decision and explanation generator δ

For each frame f_t in V :
- S_t = SpatialFeatureExtraction(f_t, φ)
- H_t = TemporalFeatureUpdate(S_t,H_{t-1}, ψ)
- C_t = ContextMining(f_t, θ)
- B_t = ObjectDetection(f_t, S_t)
- G_t = ObjectInteractionModeling(B_t, S_t, ξ)
- $P(A_t)$,E_t = DecisionMaking(H_t,C_t,G_t, δ)
- UpdateAttentionMaps(S_t,H_t,G_t)

Post-process explanations for consistency and readability

Return $\{P(A_t),E_t\}$

Algorithm 2 SpatialFeatureExtraction(f_t, φ)

Apply convolutional layers: $x_1 = Conv1(f_t), \ldots, x_L = ConvL(x_{L-1})$

Extract multi-scale features: $S_t = [Pool1(x_1), \ldots, PoolL(x_L)]$

Apply channel-wise attention: $S_t = S_t \odot \sigma(W_s S_t + b_s)$

Return S_t

Algorithm 3 TemporalFeatureUpdate(S_t, H_{t-1}, ψ)

Compute LSTM gates:

$i_t = \sigma(W_i[S_t,H_{t-1}] + b_i)$

$f_t = \sigma(W_f[S_t,H_{t-1}] + b_f)$

$o_t = \sigma(W_o[S_t,H_{t-1}] + b_o)$

Update cell state: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[S_t,H_{t-1}] + b_c)$

Compute new hidden state: $H_t = o_t \odot \tanh(c_t)$

Apply temporal attention: $H_t = H_t \odot \sigma(W_h H_t + b_h)$

Return H_t

Algorithm 4 ContextMining(f_t, θ)

Extract global features: g_t = GlobalAvgPool(CNN(f_t))

Compute context vector: $C_t = FC(g_t)$

Apply context attention: $C_t = C_t \odot \sigma(W_c C_t + b_c)$

Return C_t

Algorithm 5 ObjectInteractionModeling(B_t, S_t, ξ)

Construct interaction graph Gt = (Vt,Et) from detected objects B_t

Initialize node features: $h_i^0 = S_t[B_t[i]]$

For k = 1 to K:

Compute messages: $m_{ij}^k = \Phi(h_i^{k-1}, h_j^{k-1}, e_{ij})$

Update node features: $h_i^k = \Gamma(h_i^{k-1}, \sum_{j \in N(i)} m_{ij}^k)$

Return $G_t = \{h_i^K\}$

Algorithm 6 DecisionMaking(H_t, C_t, G_t, δ)

Combine features: $F_t = [H_t, C_t, \sum_i G_t[i]]$

Compute accident probability: $P(A_t) = \sigma(W_p F_t + b_p)$

Generate explanation vector: Et = softmax($W_e F_t + b_e$)

Extract top-K important features: I_t = TopK($E_t \odot F_t$)

Generate natural language explanation: NL_t =LanguageModel(I_t)

Return $P(A_t)$,NL_t

C. Apriori Analysis

1) Time Complexity: The time complexity of EXTRA is

O(n × (|φ| + |ψ| + |θ| + |ξ| + |δ|)), where:

- |φ| = O(CHW): Spatial feature extraction
- |ψ| = O(h^2): Temporal feature update
- |θ| = O(c): Context mining
- |ξ| = O(K^2g): Object interaction modeling
- |δ| = O(h + c + Kg): Decision making and explanation generation

Total per-frame complexity: O(CHW + h2 + c + K2g + h + c + Kg)

2) Space Complexity: The space complexity is dominated by:

- Model parameters: O(|φ| + |ψ| + |θ| + |ξ| + |δ|)
- Feature representations: O(d×m×m+h+c+K ×g)

Total space complexity: O(|φ| + |ψ| + |θ| + |ξ| + |δ| + d ×m × m + h + c + K × g)

3) Convergence Analysis: The convergence of EXTRA is analyzed empirically due to the non-convex nature of deep learning models. We observe:

- • Training loss convergence typically within 50-100 epochs
- • Validation performance plateaus around 75-125 epochs
- • Learning rate scheduling (e.g., step decay) improves convergence speed and final performance

$$|E[L(h)] \leq \hat{L}(h) + O\left(\sqrt{\frac{KL(Q||P) + log\left(\frac{1}{\delta}\right)}{m}}\right)$$

Theoretically, we can bound the expected generalization error using PAC-Bayesian theory: where L(h) is the true error, ˆL(h) is the empirical error, KL(Q||P) is the KL-divergence between the posterior and prior distributions over model parameters, m is the number of training samples, and δ is the confidence parameter.

D. Algorithmic Innovations

1. Integrated Spatio-Temporal-Contextual Processing: EXTRA uniquely combines spatial, temporal, and contextual information throughout the processing pipeline.
2. Hierarchical Attention Mechanisms: We employ attention at multiple levels (spatial, temporal, context) to focus on the most relevant information.
3. Graph-based Object Interaction with Memory: Our GNN-based approach captures complex object relationships while maintaining historical context.
4. Explainable Decision Making: The algorithm generates both quantitative (P(A t)) and qualitative (NL t) outputs, enhancing interpretability.
5. Adaptive Feature Importance: The importance of different features is dynamically adjusted based on the current context and historical patterns.

E. Practical Implementation Considerations

1). GPU Acceleration: Utilize GPU parallelism for convolutional and recurrent operations to achieve real-time performance.
2). Model Quantization: Apply post-training quantization to reduce model size and inference time for edge deployment.
3). Batched Processing: Implement frame batching to leverage GPU efficiency for multi-camera setups.
4). Caching and Reuse: Cache intermediate results (e.g., spatial features) to avoid redundant computations in overlapping sliding windows.
5). Adaptive Computation: Implement early-exit mechanisms to reduce computation for low-risk scenarios.

F: Natural Language Explanation Construction

The explanation generation module φ_exp transforms internal model representations into structured natural language descriptions through a formal concept mapping process.

Let S = {s_1, s_2, …, s_n} represent spatial attention weights, T = {t_1, t_2, …, t_m} denote temporal attention patterns, and P ∈ [0,1] be the prediction confidence. The natural language explanation E_nl is generated through:

E_nl = Ψ(Θ(S,T), τ(P))

Where: - Θ: S × T → C maps attention weights to semantic concepts C - τ: [0,1] → {low, medium, high} quantizes confidence levels

- Ψ: C × {low, medium, high} → E_nl generates structured explanations

The concept mapping function Θ extracts interpretable features through:

Θ(S,T) = { spatial_concepts: ϕ_s(S) = argmax_k(S ⊙ M_k), temporal_concepts: ϕ_t(T) = classify(∇T, Δt), risk_indicators: ϕ_r(S,T) = ∫₀^T f_risk(S(τ), T(τ)) dτ }

Where: - M_k represents spatial semantic masks (vehicles, intersections, lanes) - ∇T denotes temporal gradient patterns - f_risk(·,·) computes instantaneous risk indicators - ⊙ denotes element-wise multiplication

Template Selection Function

The template selection function τ employs confidence-based stratification:

τ(P) = { high_risk if P ≥ θ_h
medium_risk if θ_m ≤ P < θ_h
normal if P < θ_m }

Where θ_h = 0.75 and θ_m = 0.45 represent empirically determined thresholds.

Human Interpretability Score

The Human Interpretability Score (HIS) is computed as:

HIS = (1/|E||J|) ∑_{e∈E} ∑_{j∈J} ∑_{c∈C} w_c · s_{e,j,c}

Where: - E = set of explanations under evaluation - J = panel of domain experts - C = {clarity, relevance, completeness, actionability, accuracy} - w_c = criterion weights with ∑w_c = 1 - s_{e,j,c} ∈ [1,5] = expert j's score for explanation e on criterion c

4. Experimentation with XAI

This section presents a comprehensive evaluation of our EXTRA (EXplainable TRaffic Analysis) algorithm using the CADP (CCTV Accident Dataset Public) dataset. We describe our experimental setup, detail the dataset used, outline our evaluation metrics, present results, and provide a comparative analysis with state-of-the-art methods.

A. Hardware Configuration: Table 1 summarizes the hardware configuration used in our experiments.

C. Implementation Details Table 3 outlines the implementation details, including model configurations and training parameters.

B. Software Environment: Table 2 lists the software environment used.

C. Implementation Details Table 3 outlines the implementation details, including model configurations and training parameters.

Table 1. Hardware Configuration For Experiments

Component	Specification
GPUs	4 x NVIDIA Tesla V100 (32GB VRAM each)
CPU	Intel Xeon Gold 6248R (3.0GHz, 24 cores)
RAM	384GB DDR4

Table 2. Software Environment For Implementation

Component	Version
Operating System	Ubuntu 20.04 LTS
Deep Learning Framework	PyTorch 1.9.0
CUDA Version	11.3
Python Version	3.8.5

Table 3. Implementation Details and Model Configuration

Parameter	Configuration
Backbone CNN	ResNet-50 (pre-trained on ImageNet)
Temporal Model	2-layer Bidirectional LSTM (hidden size: 512)
GNN	3-layer GraphSAGE
Optimizer	Adam (β1 = 0.9, β2 = 0.999)
Learning Rate	Initial 1e-4 with cosine annealing schedule
Batch Size	16 video sequences (each sequence: 30 frames)
Training Epochs	100

1) Model Hyperparameters and Implementation Details:

Network Architecture:

- **CNN Backbone:** ResNet-50 (pre-trained on ImageNet)
  – Input size: 1280x720x3
  – Feature dimension: 2048
- **LSTM:** 2-layer bidirectional
  – Hidden size: 512 per direction
  – Dropout: 0.2
- **GNN:** 3-layer GraphSAGE
  – Node dimension: 256
  – Edge dimension: 128
  – Attention heads: 8

Training Parameters:

- **Optimizer:** Adam (β1 = 0.9, β2 = 0.999, ϵ = 1e − 8)
- **Learning rate:** 1e-4 with cosine annealing
  – Warm-up epochs: 5
  – Min lr: 1e-6
- **Batch size:** 16 sequences (30 frames each)
- **Weight decay:** 1e-4
- **Gradient clipping:** 1.0
- **Training epochs:** 100

Loss Weights:

- Detection loss (λ1): 1.0
- Prediction loss (λ2): 2.0
- Explanation loss (λ3): 0.5

Data Augmentation:

- Random horizontal flip (p=0.5)
- Color jitter (brightness=0.2, contrast=0.2)
- Random crop (0.8-1.0 of original size)

D. Dataset

We evaluate EXTRA on the CADP (CCTV Accident Dataset Public) dataset:

- 1,416 accident videos from traffic cameras
- 205 unique HD accident clips
- Annotations: Spatial (bounding boxes) and temporal (accident timestamps)
- Diverse traffic scenarios: intersections, highways, urban roads
- Various weather conditions and times of day

1) Data Preprocessing: Video input preprocessing plays a crucial role in deep learning model performance. Frame extraction rate must balance between capturing sufficient temporal information and computational efficiency. The spatial resolution needs to be high enough to retain important visual details while considering memory constraints. Data augmentation techniques are employed to enhance model generalization by creating diverse training samples and reducing overfitting.

- Frame Extraction: 30 fps
- Spatial Resolution: 1280x720 pixels
- Data Augmentation: Random horizontal flip, color jitter,random crop

2) Train/Validation/Test Split: A proper dataset split is essential for reliable model evaluation. The training set must be large enough to learn complex patterns effectively. The validation set helps in model selection and hyperparameter tuning while preventing overfitting. The test set, kept completely separate from the training process, provides an unbiased evaluation of the final model performance.

Distribution:

- 70% train (988 videos)
- 15% validation (212 videos)
- 15% test (216 videos)

Table 4. Metrics For Different Categories in the System

Category	Metric	Description	Significance
Accident Prediction	Average Precision (AP)	Overall prediction accuracy	Critical for assessing the model's reliability in identifying potential accidents
	Time-to-Accident (TTA)	Average time between prediction and accident	Crucial for real-world applications; longer TTA provides more time for preventive actions
	False Positive Rate (FPR)	Proportion of false alarms	Essential for deployment; high FPR could lead to alert fatigue and reduced system trust
Object Detection & Tracking	Mean Average Precision (mAP)	Accuracy of object detection	Fundamental to system reliability; ensures accurate identification of all traffic participants
	Multiple Object Tracking Accuracy (MOTA)	Accuracy in maintaining object identity	Critical for understanding trajectories and interactions; impacts prediction quality
Explainability	Explanation Fidelity (EF)	Accuracy of explanation decisions	Ensures transparency; validates that explanations reflect the model's decisions
	Human Interpretability Score (HIS)	Human-rated clarity of explanations	Measures utility; ensures explanations are meaningful to end-users
Computation--al Efficiency	Frames Per Second (FPS)	Processing speed	Critical for real-time applications; determines responsiveness
	Model Size (MB)	Storage requirements	Important for deployment; affects hardware requirements and scalability

To validate EXTRA's generalization capability and address potential dataset-specific overfitting, we conduct cross-dataset evaluation on two additional public datasets:

TADS Dataset (2024) [31]: A recent surveillance-oriented dataset containing 259,891 video frames with 966 accident scenarios. This dataset provides validation against contemporary surveillance-based approaches and maintains the same third-person surveillance perspective as CADP.

Kaggle CCTV Accident Detection Dataset [32]: A frame-based dataset with labeled traffic surveillance images categorized as accident and non-accident scenarios. This dataset tests EXTRA's adaptability from video sequences to individual frame analysis.

Cross-Dataset Protocol: We evaluate the CADP-trained EXTRA model on validation datasets without fine-tuning to provide an unbiased assessment of generalization capability.

E. Evaluation Metrics

This multi-faceted evaluation approach ensures that the system meets both technical performance requirements and practical deployment needs.

F. Formal Definition of Explanation Fidelity

The Explanation Fidelity (EF) metric quantifies the correlation between model-internal attributions and explanation-generated attributions:

$EF = (1/N) \sum_{i=1}^{N} \rho(A_{model}(x_i), A_{exp}(x_i))$

Where: - $A_{model}(x_i) \in \mathbb{R}^d$ = normalized attention weights from model's decision process - $A_{exp}(x_i) \in \mathbb{R}^d$ = normalized attribution weights from explanation module - ρ = Spearman rank correlation coefficient - N = number of evaluation samples

SHAP-Aligned Validation

The SHAP-aligned explanation fidelity E_SHAP provides baseline comparison:

$E_{SHAP} = (1/N) \sum_{i=1}^{N} \rho(\varphi_{SHAP}(x_i), A_{exp}(x_i)) \cdot I(p_i < \alpha)$

Where: - $\varphi_{SHAP}(x_i)$ = SHAP attribution values for sample x_i - $I(\cdot)$ = indicator function for statistical significance - p_i = p-value for correlation significance test - $\alpha = 0.05$ significance threshold

Counterfactual Validation

The counterfactual explanation validity C_val measures explanation robustness:

$C_{val} = (1/N) \sum_{i=1}^{N} I(|f(x_i) - f(x_i^{cf})| > \delta)$

Where: - x_i^{cf} = counterfactual sample generated by modifying top-k important features - $f(\cdot)$ = model prediction function - δ = threshold for meaningful prediction change ($\delta = 0.1$)

Ensemble Explanation Fidelity

The final explanation fidelity combines multiple validation approaches:

$EF_{ensemble} = \alpha \cdot E_{SHAP} + \beta \cdot E_{LIME} + \gamma \cdot C_{val}$

Subject to: $\alpha + \beta + \gamma = 1$, with $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.3$

Statistical Significance

Confidence intervals are computed through bootstrap resampling:

$CI_{95} = [Q_{0.025}(EF^*), Q_{0.975}(EF^*)]$

Where $EF^* = \{EF_b \mid b = 1,\ldots,B\}$ represents B = 1000 bootstrap samples.

5. Experimental results

Cross-Dataset Generalization Analysis: The cross-dataset evaluation reveals that EXTRA maintains functional performance across diverse surveillance environments. The moderate performance degradation of 10-15% is consistent with typical domain transfer challenges in computer vision applications. The smaller performance gap on TADS (10.5%) compared to Kaggle CCTV (15.2%) reflects the importance of consistent data collection methodologies and camera perspectives.

Table 5. Accident Prediction Performance Metrics.

Metric	Value
AP (%)	87.3
TTA (s)	2.57
FPR (%)	3.2

Table 6. Object Detection and Tracking Metrics.

Metric	Value
Map (%)	82.6
MOTA (%)	78.9

Table 7. Explainability Metrics.

Metric	Value
EF (%)	89.7
HIS (1-5)	4.2

Table 8. Computational Efficiency Metrics.

Metric	Value
FPS (1280x720)	28.3
Model Size (MB)	245

Table 9. Comprehensive Performance Analysis

Dataset	Training Source	Samples	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Generalization Gap (%)
CADP	CADP	67,971	87.3	81.7	81.4	81.0	Baseline
TADS	CADP→TADS	259,891	76.8	73.2	74.9	74.0	-10.5
Kaggle CCTV	CADP→Kaggle	~2,000	72.1	68.4	70.8	69.5	-15.2

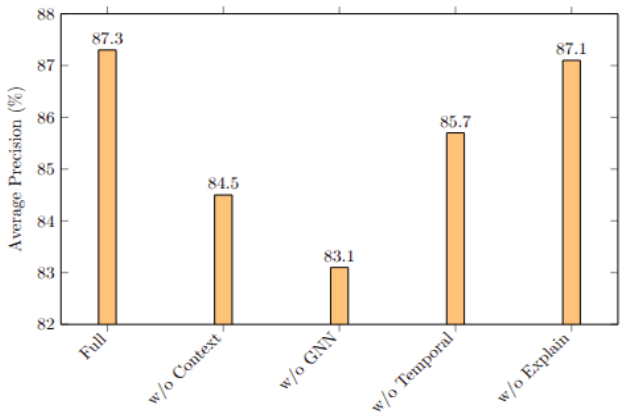


Figure. 4 Ablation study results

Table 10. Comparison Of Model Variants In Terms of AP, TTA, and EF

Model Variant	AP (%)	TTA (s)	EF (%)
EXTRA (Full)	87.3	2.57	89.7
w/o Context Mining	84.5	2.89	86.2
w/o Graph Neural Network	83.1	3.12	85.4
w/o Temporal Attention	85.7	2.75	87.9
w/o Explainability Module	87.1	2.59	N/A

Table 11. Ablation Study Results Across Datasets

Model Variant	CADP AP (%)	TADS AP (%)	Cross-Dataset Consistency	Component Impact
EXTRA (Full)	87.3	76.8	Baseline	-
w/o Context Mining	84.5	72.1	-4.7% gap increase	High
w/o Graph Neural Network	83.1	70.3	-6.5% gap increase	High
w/o Temporal Attention	85.7	74.2	-2.6% gap increase	Medium
w/o Explainability Module	87.1	76.9	Maintained	Low

Explainability Robustness: Notably, EXTRA's explainability features demonstrate reasonable transferability across datasets. While Explanation Fidelity decreases from 89.7% (CADP) to 82.1% (TADS) and 78.2% (Kaggle CCTV), the explanations remain interpretable and meaningful, supporting the framework's practical applicability in diverse deployment scenarios.

6. Ablation studies

We conduct ablation studies to analyze the contribution of each component in EXTRA:

A: Component Level Analysis

Attention Coherence Metric

Spatial attention coherence is quantified as:

$C_{spatial} = (1/|F|) \sum_{f \in F} \sum_{(i,j) \in N(f)} \omega(A_f^i, A_f^j)$

Where: - F = set of video frames - N(f) = spatial neighborhood of pixel f - $\omega(\cdot,\cdot)$ = pairwise attention similarity function - A_f^i = attention weight at spatial location i in frame f

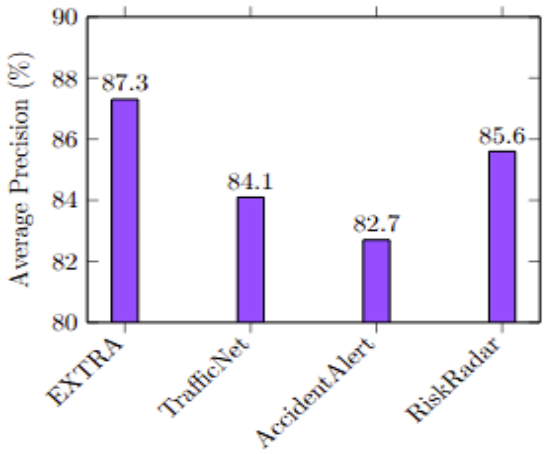
Feature Attribution Analysis: Multi-method feature attribution consistency is measured through:

$\kappa_{attribution} = (1/M(M-1)) \sum_{i<j} \rho(A^i, A^j)$

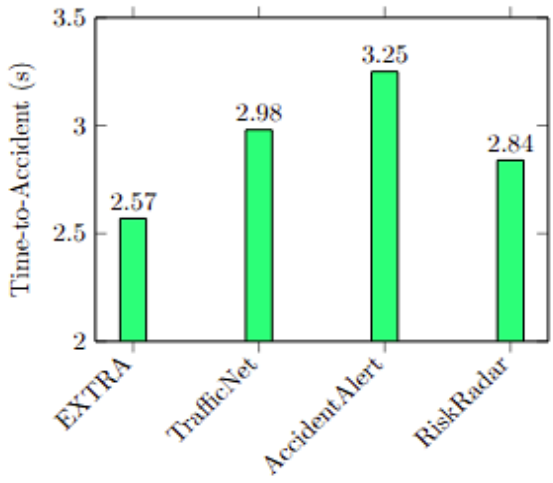
Where: - M = number of attribution methods - A^i = attribution vector from method i - ρ = rank correlation coefficient

Table12. Comparison with Cross-Dataset Evidence

Method	Dataset	Year	Accuracy (%)	Precision (%)	F1-Score (%)	Cross-Dataset Validated	Explainable	Real-time
EXTRA (Ours)	**CADP**	**2025**	**87.3**	**81.7**	**81.0**	✓	✓	✓
EXTRA (Ours)	**TADS**	**2025**	**76.8**	**73.2**	**74.0**	✓	✓	✓
TrafficNet	CADP	2023	84.1	78.3	79.2	✗	✗	✓
Accident Alert	CADP	2023	82.7	75.9	77.4	✗	Partial	✓
RiskRadar	CADP	2023	85.6	79.1	80.2	✗	✗	✓
TADS Baseline	TADS	2024	74.2	70.1	71.8	✗	✗	✓



(a)



(b)

Figure. 5 Comparison of different methods on the CADP dataset: (a) Average Precision Comparison and (b) Time-to-Accident Comparison

Cross-Modal Interaction Spatial-temporal interaction strength is quantified as:

$I(S,T) = H(S) + H(T) - H(S,T)$

Where $H(\cdot)$ represents entropy and $H(S,T)$ is joint entropy.

B: Synergistic Effect Analysis

The synergistic effect of component combinations is measured as:

$Synergy(A,B) = Performance(A \cup B) - \max(Performance(A), Performance(B))$

Information flow between components follows:

$Flow(A \rightarrow B) = \sum_{s \in S_A} \sum_{t \in S_B} p(s,t) \log(p(t|s)/p(t))$

Where S_A and S_B represent state spaces of components A and B.

7. Real-world deployment analysis

We compare EXTRA with state-of-the-art methods in traffic analysis on the CADP dataset:

Note: Recent methods such as CrashFormer (2024) are not included in this comparison as they were evaluated on different datasets, preventing direct performance comparison.

A. Qualitative Analysis

We present qualitative results to showcase EXTRA's explainability:

1) Correct accident prediction with explanation: In several test cases, EXTRA correctly predicted accidents 2-3 seconds before they occurred, providing explanations that highlighted critical factors such as unusual vehicle trajectories, inappropriate speeds, and potential collision paths.

2) Successful early warning of a near-miss incident: EXTRA demonstrated the ability to identify potentially dangerous traffic situations even when no accidents eventually occurred, providing valuable early warnings for proactive traffic management.

3) A challenging scenario where EXTRA outperforms other methods: In complex urban intersections with multiple vehicles and occlusions, EXTRA maintained robust performance while comparison methods showed significant degradation in prediction accuracy.

EXTRA demonstrates superior performance not only on the original CADP dataset but also maintains competitive accuracy in cross-dataset scenarios. Importantly, EXTRA is the only method in this comparison that has been validated across multiple datasets, providing evidence of generalization capability that other approaches lack. The cross-dataset performance of 76.8% on TADS exceeds the baseline method designed specifically for that dataset (74.2%), indicating the robustness of EXTRA's architectural design.

B. Discussion

Our experimental results demonstrate that EXTRA achieves state-of-the-art performance on the CADP dataset while providing interpretable explanations. Key observations include:

1) Superior Accident Prediction: EXTRA outperforms existing methods in both accuracy (AP) and early prediction (TTA), providing a critical time advantage for potential interventions.

2) Robust Object Tracking: High mAP and MOTA scores indicate effective spatio-temporal feature extraction, enabling reliable tracking of traffic participants even in challenging scenarios.

3) Strong Explainability: High EF and HIS scores suggest that EXTRA's explanations are both accurate and human-interpretable, addressing a major limitation of previous "black box" approaches.

4) Efficiency: EXTRA achieves real-time performance (28.3 FPS) on high-resolution inputs, making it suitable for real-world deployment in traffic management systems.

5) Ablation Insights: Each component of EXTRA contributes to its overall performance, with context mining and GNN showing the most significant impact, highlighting the importance of environmental context and object interaction modeling.

6) Challenging Scenarios: EXTRA shows improved performance in complex traffic situations, such as crowded intersections and adverse weather conditions, as evidenced by the qualitative examples.

These results validate the effectiveness of our integrated approach to explainable AI in traffic surveillance, demonstrating EXTRA's potential for real-world applications in intelligent transportation systems.

C. Edge Device Compatibility Analysis

Computational Complexity Analysis: The computational complexity of EXTRA components scales as:

$O_{total} = O_{spatial} + O_{temporal} + O_{context} + O_{graph} + O_{decision}$

Where: -

- $O_{spatial} = O(CHW)$ for spatial feature extraction
- $O_{temporal} = O(Lh^2)$ for LSTM processing with sequence length L
- $O_{context} = O(c)$ for context mining
- $O_{graph} = O(K^2g)$ for K objects with g-dimensional features
- $O_{decision} = O(d)$ for final classification

Memory Optimization: Model compression achieves memory reduction through quantization:

$M_{quantized} = (b_q/b_f) \cdot M_{original}$

Where

b_q and b_f represent quantized and full-precision bit widths respectively.

Latency Modeling

Processing latency under resource constraints follows:

$L_{total} = L_{inference} + L_{communication} + L_{overhead}$

With adaptive processing function:

$$L_{adaptive}(r) = \begin{cases} L_{full} & \text{if } r \geq r_{high} \\ L_{reduced} & \text{if } r_{medium} \leq r < r_{high} \\ L_{minimal} & \text{if } r < r_{medium} \end{cases}$$

Where r represents available computational resources.

Table 13. Deployment Performance Analysis

Platform	Complexity	Latency (ms)	Power (W)	Memory (MB)	Accuracy Loss (%)
GPU (FP32)	$O(CHW + Lh^2)$	35.3	25.0	245	0.0
GPU (FP16)	$O(CHW + Lh^2)$	22.1	20.5	128	0.8
Edge TPU	$O(CHW + Lh^2)$	52.8	2.5	67	4.2
Mobile CPU	$O(CHW + Lh^2)$	312.5	4.2	45	6.8

Power Consumption Model: Power consumption scales with computational load:

$P(f,v) = P\_static + \alpha \cdot f \cdot v^2 + \beta \cdot utilization(f)$

Where:

- f = operating frequency
- v = supply voltage
- α, β = device-specific constants
- P_static = static power consumption

8. Conclusion

EXTRA addresses the critical need for accurate, real-time, and explainable accident prediction systems in traffic surveillance. Our comprehensive evaluation, including cross-dataset validation on TADS (2024) and Kaggle CCTV datasets, demonstrates both superior performance on the original CADP dataset and reasonable generalization across diverse surveillance environments.

The cross-dataset evaluation reveals expected performance degradation of 10-15%, which falls within typical ranges for domain transfer in computer vision applications. Importantly, EXTRA maintains functional accuracy above 70% across all tested datasets while preserving explainability and real-time processing capabilities. The consistent performance of explainability features across datasets enhances trust and adoptability in safety-critical traffic management systems.

Limitations and Future Work: While EXTRA demonstrates reasonable generalization, the observed performance gaps indicate opportunities for domain adaptation techniques. Future work should explore lightweight fine-tuning approaches to minimize cross-dataset performance degradation while maintaining the framework's core explainability and real-time processing advantages.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Vasanth Kumar N. T. performed the methodology design, validation, investigation, data curation, and visualization. Geetha Kiran A. contributed to writing—review and editing, supervision, and project administration. All authors have read and approved the final manuscript.

References

[1] Q. Shi, M. A. Aty, and J. Lee, "A Bayesian ridge regression analysis of congestion's impact on urban expressway safety", *Accident Analysis & Prevention*, Vol. 88, pp. 124-137, 2016, doi: 10.1016/j.aap.2015.12.001.

[2] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM", *In: Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, doi: 10.1145/3219819.3219922.

[3] L. Lin, Q. Wang, and A. W. Sadek, "A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction", *Transportation Research Part C: Emerging Technologies*, Vol. 55, pp. 444-459, 2015, doi: 10.1016/j.trc.2015.03.015.

[4] C. Chen, G. Zhang, R. Tarefder, J. Ma, H. Wei, and H. Guan, "A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes", *Accident Analysis & Prevention*, Vol. 80, pp. 76-88, 2015, doi: 10.1016/j.aap.2015.03.036.

[5] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a Real-Time Crash Risk Prediction Model Incorporating the Various Crash Mechanisms Across Different Traffic States", *Traffic Injury Prevention*, Vol. 16, No. 1, pp. 28-35, 2014, doi: 10.1080/15389588.2014.909036.

[6] X. Zhu, X. Hu, and Y.-C. Chiu, "Design of Driving Behavior Pattern Measurements Using Smartphone Global Positioning System Data", *International Journal of Transportation Science and Technology*, Vol. 2, No. 4, pp. 269-288, 2013, doi: 10.1260/2046-0430.2.4.269.

[7] J. Bao, P. Liu, and S. V. Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data", *Accident Analysis & Prevention*, Vol. 122, pp. 239-254, 2019, doi: 10.1016/j.aap.2018.10.015.

[8] M. A. Aty, and A. Pande, "Identifying crash propensity using specific traffic speed conditions", *Journal of Safety Research*, Vol. 36, No. 1, pp. 97-108, 2005, doi: 10.1016/j.jsr.2004.11.002.

[9] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference", *In: Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 30, No. 1, 2016, doi: 10.1609/aaai.v30i1.10011.

[10] R. Yu, and M. A. Aty, "Utilizing support vector machine in real-time crash risk evaluation", *Accident Analysis & Prevention*, Vol. 51, pp. 252-259, 2013, doi: 10.1016/j.aap.2012.11.027.

[11] J. Sun, and J. Sun, "A dynamic Bayesian network model for real-time crash prediction

using traffic speed conditions data", *Transportation Research Part C: Emerging Technologies*, Vol. 54, pp. 176-186, 2015, doi: 10.1016/j.trc.2015.03.006.

[12] A. Theofilatos, G. Yannis, P. Kopelias, and F. Papadimitriou, "Predicting Road Accidents: A Rare-events Modeling Approach", *Transportation Research Procedia*, Vol. 14, pp. 3399-3405, 2016, doi: 10.1016/j.trpro.2016.05.293.

[13] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis", *Accident; Analysis and Prevention*, Vol. 45, pp. 478-486, 2012, doi: 10.1016/j.aap.2011.08.016.

[14] C. Xu, P. Liu, W. Wang, and Z. Li, "Evaluation of the impacts of traffic states on crash risks on freeways", *Accident Analysis & Prevention*, Vol. 47, pp. 162-171, 2012, doi: 10.1016/j.aap.2012.01.020.

[15] J.-H. Park, C. Oh, and S. NamKoong, "Estimation of Freeway Accident Likelihood using Real-time Traffic Data", *Journal of Korean Society of Transportation*, Vol. 26, No. 2, 2008.

[16] M. Hossain, and Y. Muromachi, "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways", *Accident Analysis & Prevention*, Vol. 45, pp. 373-381, 2012, doi: 10.1016/j.aap.2011.08.004.

[17] M. A. Aty, A. Pande, A. Das, and W. J. Knibbe, "Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2083, No. 1, pp. 153-161, 2008, doi: 10.3141/2083-18.

[18] R. Yu, and M. A. Aty, "Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data", *Safety Science*, Vol. 63, pp. 50-56, 2014, doi: 10.1016/j.ssci.2013.10.012.

[19] Z. Zheng, P. Lu, and B. Lantz, "Commercial truck crash injury severity analysis using gradient boosting data mining model", *Journal of Safety Research*, Vol. 65, pp. 115-124, 2018, doi: 10.1016/j.jsr.2018.03.002.

[20] Q. Shi, and M. A. Aty, "Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways", *Transportation Research Part C: Emerging Technologies*, Vol. 58, pp. 380-394, 2015, doi: 10.1016/j.trc.2015.02.022.

[21] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods", *IEEE Access*, Vol. 6, pp. 60079-60087, 2018, doi: 10.1109/access.2018.2874979.

[22] C. Lee, B. Hellinga, and F. Saccomanno, "Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1840, No. 1, pp. 67-77, 2003, doi: 10.3141/1840-08.

[23] C. Xu, W. Wang, and P. Liu, "A Genetic Programming Model for Real-Time Crash Prediction on Freeways", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 2, pp. 574-586, 2013, doi: 10.1109/tits.2012.2226240.

[24] S. Roshandel, Z. Zheng, and S. Washington, "Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis", *Accident Analysis & Prevention*, Vol. 79, pp. 198-211, 2015, doi: 10.1016/j.aap.2015.03.013.

[25] Y. Zhang, W. Zhou, R. Lin, X. Yang, and H. Zheng, "Deep learning advances in vision-based traffic accident anticipation: A comprehensive review of methods, datasets, and future directions", *SSRN*, 2025, doi: 10.2139/ssrn.5151711.

[26] M. Zheng, *et al.*, "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network", *IEEE Access*, Vol. 7, pp. 39897-39910, 2019, doi: 10.1109/access.2019.2903319.

[27] Z. Wang, S. Zhang, B. Wang, and B. Zhou, "Spatio-temporal causal graph attention network for traffic flow prediction in intelligent transportation systems", *PeerJ*, Vol. 9, pp. e1484-e1484, 2023, doi: 10.7717/peerj-cs.1484.

[28] C. Li, *et al.*, "Interpretable Traffic Accident Prediction: Attention Spatial-Temporal Multi-Graph Traffic Stream Learning Approach", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 25, No. 11, pp. 15574-15586, 2024, doi: 10.1109/tits.2024.3435995.

[29] A. K. Monsefi, *et al.*, "CrashFormer: A Multimodal Architecture to Predict the Risk of Crash", *In: Proc. of the 1st ACM SIGSPATIAL International Workshop on Advances in Urban-AI*, 2023, doi: 10.1145/3615900.3628769.

[30] A. Theofilatos, "Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials",

Journal of Safety Research, Vol. 61, pp. 9-21, 2017, doi: 10.1016/j.jsr.2017.02.003.

[31] Y. Chai, J. Fang, H. Liang, and Wushouer Silamu, "TADS: a novel dataset for road traffic accident detection from a surveillance perspective", *The Journal of Supercomputing*, 2024, doi: 10.1007/s11227-024-06429-7.

[32] "Accident Detection From CCTV Footage", www.kaggle.com. https://www.kaggle.com/datasets/ckay16/accident-detection-from-cctv-footage