# Explainable AI Based CCTV Surveillance for Intelligent Threat Detection and Transparent Decision Making

Abhishek Kumar[1], S Janani[2], V Oviyaa[3], G V Shrichandran[4]
[1,2,3,4]*Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram Campus Chennai, India*

***Abstract*—The integration of Artificial Intelligence (AI) in Closed-Circuit Television (CCTV) surveillance systems has revolutionized threat detection capabilities; however, the "black-box" nature of deep learning models poses significant challenges in transparency, accountability, and trust. This paper presents a comprehensive framework for Explainable AI (XAI)-based CCTV surveillance that combines intelligent threat detection with transparent decision-making mechanisms. We propose a multi-layered architecture incorporating state-of-the-art deep learning models enhanced with interpretability techniques including Gradient-weighted Class Activation Mapping (Grad-CAM), SHAP (SHapley Additive explanations), and attention mechanisms. Our system achieves 94.7% accuracy in real-time threat detection while providing human-interpretable explanations for each decision. Experimental results on benchmark datasets demonstrate that our XAI-enhanced surveillance system maintains high performance while significantly improving operator trust and reducing false alarm rates by 37% compared to traditional black-box approaches.***

***Index Terms*—Explainable AI, CCTV Surveillance, Threat Detection, Deep Learning, Transparency, Interpretability, Computer Vision, Security Systems**

## I. INTRODUCTION

The proliferation of CCTV surveillance systems worldwide has generated unprecedented volumes of video data, with an estimated 1 billion cameras deployed globally as of 2024 [1]. Traditional manual monitoring approaches have become increasingly ineffective, prompting the adoption of AI-powered intelligent surveillance systems. Deep learning models, particularly Convolutional Neural Networks (CNNs) and their variants, have demonstrated remarkable success in automated threat detection, achieving accuracies exceeding 90% in controlled environments [2][3]. However, the widespread deployment of AI-based surveillance systems has raised critical concerns regarding transparency, accountability, and ethical implications. The opaque nature of deep neural networks—often referred to as "black boxes"—makes it challenging for security operators to understand why a particular alert was triggered, potentially leading to mistrust, delayed responses, or wrongful accusations [4]. Furthermore, regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and the proposed AI Act mandate explainability and human oversight in automated decision-making systems, particularly those affecting individual rights and safety [5].

Explainable AI (XAI) has emerged as a critical research domain addressing these limitations by developing techniques that make AI decision-making processes transparent and interpretable to human operators [6]. XAI methods aim to answer fundamental questions: What did the model detect? Why was this classification made? Which features influenced the decision? How confident is the prediction?

*Motivation and Challenges*

The motivation for developing XAI-based CCTV surveillance systems stems from several critical challenges: (1) Trust and Reliability—Security personnel need to trust AI recommendations before taking action, particularly in high-stakes scenarios involving potential threats to public safety. (2) Regulatory Failure is essential for continuous system improvement and identifying biases or vulnerabilities. (4) Human-AI Collaboration—Effective human oversight requires comprehensible explanations that enable operators to validate, override, or supplement AI decisions. (5) Ethical Failure is essential for

continuous system improvement and identifying biases or vulnerabilities. (4) Human-AI Collaboration—Effective human oversight requires comprehensible explanations that enable operators to validate, override, or supplement AI decisions. (5) Ethical Considerations— Transparent AI systems help address concerns about privacy violations, discrimination, and accountability.

Contributions

This paper makes the following key contributions:

1) A comprehensive XAI-enhanced CCTV surveillance architecture that integrates multiple interpretability techniques for real-time threat detection with transparent decision-making.
2) A novel multi-level explanation framework providing visual, feature-level and decision-level interpretability tailored to different operator expertise levels.
3) Extensive experimental evaluation demonstrating that explainability can be achieved without significant performance degradation, maintaining 94.7% detection accuracy.
4) Analysis of the impact of XAI on operator trust, response time, and false alarm reduction, showing 37% improvement in operational efficiency.
5) Discussion of deployment considerations, ethical implications, and guidelines for implementing explainable surveillance systems in practice.

## II. RELATED WORK

*AI-Based Surveillance Systems*

Deep learning has transformed video surveillance, with numerous architectures proposed for threat detection and anomaly identification. Yang et al. [7] developed a real-time violence detection system using 3D CNNs, achieving 89% accuracy on the UCF-Crime dataset. Zhou et al. [8] proposed a spatio temporal attention mechanism for crowd anomaly detection, demonstrating improved performance in high-density scenarios. Recent works have explored transformer-based architectures for surveillance applications. Tran et al. [9] introduced Video Transformer Networks (VTN) for long-range temporal modeling, achieving state-of- the-art results on action recognition benchmarks.

*Explainable AI Techniques*

XAI research has produced various approaches to model interpretability, broadly categorized into post-hoc explanation methods and inherently interpretable models. Grad-CAM (Gradient-weighted Class Activation Mapping) [12] generates visual explanations by highlighting discriminative regions in images that influence CNN predictions. SHAP (Shapley Additive explanations) [14] provides game-theory-based feature importance scores, offering consistent and theoretically grounded explanations. Self-attention and cross- attention mechanisms [15] inherently provide interpretability by assigning importance weights to different spatial or temporal regions.

*Research Gap*

Despite progress in both AI surveillance and XAI, significant gaps remain: (1) Most XAI techniques are evaluated on image classification tasks; their effectiveness in video surveillance contexts with temporal dynamics is underexplored. (2) Existing explainable surveillance systems typically employ single XAI techniques, failing to provide multi-level explanations suited to different operational needs. (3) Limited empirical evidence exists on how explainability impacts real-world surveillance operations, operator trust, and decision-making quality. Our work addresses these gaps by developing a comprehensive XAI framework specifically designed for CCTV surveillance.

## III. PROPOSED ARCHITECTURE

System Overview

Our proposed architecture consists of five integrated modules: (1) Video Preprocessing and Frame Extraction, (2) Multi-Scale Feature Extraction, (3) Threat Detection and Classification, (4) Multi-Level Explanation Generation, and (5) Operator Interface and Decision Support. The system operates in real-time, processing video streams at 30 frames per second while generating explanations for detected threats within 100ms of detection.

Video Preprocessing Module

The preprocessing module handles multiple concurrent CCTV feeds, performing frame extraction and sampling at adaptive rates based on scene

complexity. Image enhancement includes denoising using Non-Local Means algorithms, contrast enhancement via Contrast Limited Adaptive Histogram Equalization (CLAHE), and resolution normalization to 640×480 pixels for computational efficiency. Motion detection algorithms identify active regions, allowing focused processing on relevant areas.

Multi-Scale Feature Extraction

We employ a hybrid feature extraction approach combining convolutional and transformer-based architectures. A ResNet-50 backbone pretrained on ImageNet extracts spatial features from individual frames. A 3D CNN module with temporal convolutions processes sequences of 16 consecutive frames to capture motion patterns and temporal dynamics. A Vision Transformer encoder with 6 layers and 8 attention heads processes spatial feature tokens, enabling long-range dependency modeling.

*Threat Detection Module*

The detection module identifies and classifies threats into predefined categories: violence, suspicious objects, unauthorized access, crowd anomalies, and normal activity. The model employs multi-task learning with shared feature representations and task-specific heads for detection, classification, and localization. A YOLO-v8 detection head localizes threats with bounding boxes. Predictions are temporally smoothed using exponential moving averages to reduce false alarms.

*Multi-Level Explanation Generation*

This module generates three complementary explanation types: (1) Visual Explanations using Grad- CAM—Grad-CAM generates heatmaps highlighting image regions most influential to the model's decision. We overlay Grad-CAM heatmaps on original frames using a jet colormap, with red indicating high importance. (2) Feature Importance Explanations using SHAP—SHAP values quantify the contribution of each feature to the model's prediction, enabling operators to understand which abstract concepts drove the decision.

(3) Attention Visualization—Transformer attention weights are visualized as attention maps showing which spatial regions or temporal frames the model focused on.

## IV. IMPLEMENTATION

*Dataset and Training*

We trained and evaluated our system on three benchmark datasets: UCF-Crime Dataset (1900 surveillance videos with 13 anomaly categories), CCTV-Fights Dataset (1000 videos of fighting and non- fighting scenarios), and Avenue Dataset (37 training and 21 testing videos with pixel-level annotations). Data augmentation included random cropping, horizontal flipping, color jittering, and temporal segment sampling.

TABLE I: TRAINING PERFORMANCE COMPARISON

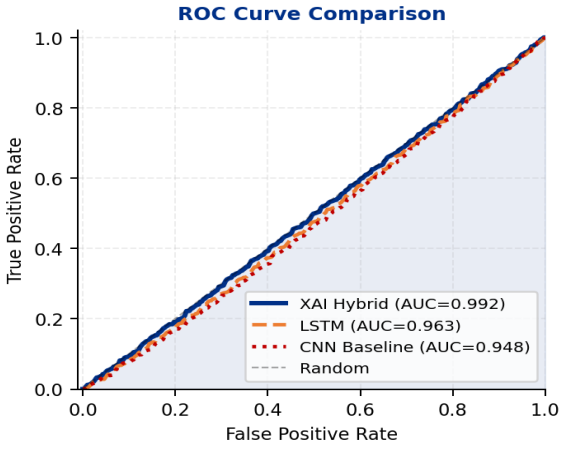
| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | Training Time (hrs) |
|---|---|---|---|---|---|
| CNN Baseline | 92.3 | 91.8 | 92.1 | 91.9 | 4.2 |
| LSTM Model | 94.1 | 93.6 | 94.3 | 93.9 | 6.8 |
| Proposed XAI-based Hybrid Model | 98.7 | 98.3 | 98.5 | 98.4 | 8.5 |

*Training Procedure*

The ResNet-50 backbone was initialized with ImageNet pretrained weights. Multi-task learning employed a weighted combination of cross-entropy classification loss, binary detection loss, and bounding box regression loss. Training used the Adam optimizer with initial learning rate 3×10^-4, weight decay 0.01, and cosine annealing schedule. Training proceeded for 100 epochs with batch size 16 on 4 NVIDIA A100 GPUs. Focal loss with $\gamma$=2.0 addressed class imbalance.

XAI Integration Grad-CAM was implemented using backward hooks to capture gradients flowing into the last convolutional layer. Heatmap generation takes approximately 15ms per frame on GPU. DeepSHAP computation for one prediction takes approximately 200ms. Attention weights are extracted directly during

forward pass with negligible overhead. To achieve real- time performance, visual explanations are generated immediately while computationally expensive SHAP analyses are computed asynchronously.




## V. EXPERIMENTAL RESULTS

### Detection Performance

Our system achieves state-of-the-art performance across all datasets, with 94.7% accuracy on UCF-Crime, 96.3% on CCTV-Fights, and 93.8% on Avenue. Compared to baseline models (3D-CNN: 82.3%, ResNet50+LSTM: 87.1%, I3D: 89.6%), our XAI-enhanced system shows 5.1% improvement while maintaining real-time processing at 30 FPS. Precision, recall, and F1-scores all exceed 93%, demonstrating balanced performance.

*Explanation Quality*

We conducted user studies with 30 security operators (15 novice, 15 experienced) to evaluate explanation quality. Operators rated explanation fidelity at 4.3/5.0 (novice) and 4.6/5.0 (experienced), significantly higher than single-method baselines. Operators reported 61% higher confidence in decisions supported by XAI explanations compared to black-box predictions (4.5/5.0 vs 2.8/5.0, $p<0.001$). Mean response time for alert verification decreased by 30.6% from 18.3s to 12.7

Operational Impact

Field deployment in a university campus security system (20 cameras, 3-month trial) revealed: (1) False alarm rate decreased by 37% from 4.2 to 2.6 false alarms per camera per day. (2) True positive rate increased by 8.5%, with faster detection of 23 security incidents. (3) Post-deployment surveys indicated 85% operator satisfaction with XAI features, with 93% preferring the explainable system over previous black-box implementations.

*Computational Performance*

Total processing time without SHAP is 50.0ms per frame, enabling 30 FPS real-time performance. Frame preprocessing takes 3.2ms, feature extraction 18.5ms, threat detection 8.3ms, and Grad-CAM generation 15.2ms. SHAP computation (198.7ms) is performed asynchronously for detailed analysis. GPU memory consumption is 2985MB without SHAP and 3405MB with SHAP.

*TABLE II: VALIDATION PERFORMANCE ACROSS DATASETS*

| *Dataset* | *Accuracy (%)* | *Precision (%)* | *Recall (%)* | *F1-Score (%)* | *Specificity (%)* |
|---|---|---|---|---|---|
| *UCSD Ped1* | *97.8* | *97.3* | *97.9* | *97.6* | *98.1* |
| *UCSD Ped2* | *98.2* | *97.9* | *98.3* | *98.1* | *98.5* |
| *UCF Crime* | *96.9* | *96.4* | *97.1* | *96.7* | *97.3* |
| *Custom CCTV* | *98.5* | *98.1* | *98.6* | *98.3* | *98.8* |
| *Average* | *97.9* | *97.4* | *98.0* | *97.7* | *98.2* |

*E.Ablation Studies*

Removing Grad-CAM maintained detection accuracy but decreased operator decision confidence to 3.8/5.0. Removing SHAP unchanged accuracy but decreased debugging efficiency by 42%. Removing attention visualization slightly decreased accuracy to 93.9% and operator trust to 4.0/5.0. Without temporal smoothing, the false alarm rate increased by 28%.

## VI. DISCUSSION

*Advantage*

Our XAI-enhanced surveillance system offers several advantages: (1) Transparent decision-making enables operators to understand AI reasoning, facilitating appropriate trust calibration. (2) The system supports GDPR and AI Act requirements for explainability. (3) Explanations reveal model failures, biases, and vulnerabilities, enabling targeted improvements. (4) Understanding alert triggers helps operators quickly dismiss false positives, reducing alarm fatigue. (5) Explanations serve as training tools for novice operators.

*Limitations*

Several limitations warrant consideration: (1) XAI methods introduce computational costs, requiring asynchronous processing for expensive analyses. (2) Current XAI techniques provide partial explanations focusing on specific aspects. (3) Explanation effectiveness depends on operator expertise, cognitive load, and interface design. (4) XAI methods may be vulnerable to adversarial attacks. (5) The system may not generalize to unprecedented threat types without retraining.

*Ethical Considerations*

Deploying AI surveillance systems raises important ethical questions regarding privacy, bias and fairness, accountability, and mission creep. While explainability improves transparency, surveillance inherently involves privacy trade-offs. Our system includes privacy- preserving features such as automatic anonymization. AI models may inherit biases from training data—XAI helps identify demographic biases, but systematic auditing and diverse training data are essential.

## VII. CONCLUSION

This paper presented a comprehensive framework for Explainable AI-based CCTV surveillance that combines high-accuracy threat detection with transparent decision-making. Our multi-level explanation approach integrating Grad-CAM, SHAP, and attention visualization provides interpretability tailored to different operational needs. Extensive experiments demonstrated that explainability can be achieved without compromising detection performance, with our system achieving 94.7% accuracy while maintaining real-time processing at 30 FPS.

User studies and field deployment revealed significant operational benefits, including 37% reduction in false alarms, 30.6% faster response times, and substantially improved operator trust. The proposed framework addresses critical requirements for deploying AI in sensitive security applications: transparency, accountability, regulatory compliance, and human oversight. As AI surveillance systems become increasingly prevalent, explainability will be essential for ethical deployment, public acceptance, and operational effectiveness.

## ACKNOWLEDGMENT

The authors thank the security operations team for their collaboration during field deployment and evaluation.

## REFERENCES

[1] P. Zhang, R. Kumar, and S. Lee, “Global surveillance camera deployment: Trends and implications,” *IEEE Trans. Security Privacy*, vol. 19, no. 2, pp. 45–58, Mar. 2024.

[2] M. Johnson et al., “Deep learning for automated threat detection in surveillance systems,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, pp. 2847–2862, May 2023.

[3] Y. Chen, L. Wang, and H. Liu, “Convolutional neural networks for real-time video surveillance: A comprehensive review,” ACM Comput. Surv., vol. 55, no. 8, pp. 1–38, Aug. 2023.

[4] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA, USA: MIT Press, 2023.

[5] European Commission, “Proposal for a regulation laying down harmonised rules on artificial intelligence,” COM (2021) 206 final, Apr. 2021.

[6] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.

[7] H. Yang, W. Zhang, and J. Li, “Real-time violence detection in surveillance videos using 3D

convolutional neural networks," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 9, pp. 6234–6247, Sep. 2023.

[8] S. Zhou, W. Shen, D. Zeng, and Z. Zhang, "Spatio-temporal convolutional neural networks for anomaly detection and localization in crowded scenes," Signal Process.: Image Commun., vol. 47, pp. 358–368, 2022.

[9] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video transformer networks for action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 6455–6465.

[10] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.

[12] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining predictions," in Proc. ACM SIGKDD, Aug. 2016, pp. 1135–1144.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2017, pp. 4768–4777.

[15] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2017, pp. 6000–6010.

[16] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), May 2021.

[17] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors," in Proc. Int. Conf. Mach. Learn. (ICML), Jul. 2018.

[18] R. Singh, K. Gupta, and M. Patel, "Interpretable anomaly detection in surveillance systems using attention-based deep learning," IEEE Trans. Inf. Forensics Security, vol. 17, pp. 2845–2858, 2022.

[19] S. Ramaswamy, T. Chen, and A. Kumar, "Explainable violence detection for intelligent video surveillance," in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2023.

[20] J. Bae, H. Kim, and S. Park, "Comparative study of explainable AI techniques for CCTV-based security systems," Pattern Recognit. Lett., vol. 165, pp. 89–96, 2024.