International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 3, May-Jun 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

IoT-Based Surveillance Systems Using Safe and Explainable AI

Mr. Balwant Singh[1], Mr. Chetan Kumar[2]

[1]M Tech Student, [2]Associate Professor,
[1,2]Kautilya Institute of Technology & Engineering, Jaipur, Rajasthan, India

ABSTRACT

This paper presents an Internet of Things (IoT)-based surveillance framework integrated with Safe and Explainable Artificial Intelligence (XAI) to resolve the transparency and reliability challenges inherent in black-box deep learning systems. Traditional autonomous surveillance models often lack interpretability, leading to high false-alarm rates, vulnerability to adversarial exploits, and delayed human verification. To address these limitations, we design a multi-sensor edge architecture that combines real-time video analytics with environmental telemetry while incorporating model-agnostic XAI frameworks (SHAP and LIME) to provide security operators with real-time, human-interpretable rationales for triggered alarms. To ensure operational safety and cyber-resilience against unauthorized intrusions, the system implements end-to-end AES-256 and SHA-512 cryptographic pipelines alongside model-pruning methodologies to eliminate edge-compute latency bottlenecks. Experimental results demonstrate that the proposed architecture achieves high threat-detection accuracy while significantly reducing Mean Time to Respond (MTTR), offering a scalable, reliable, and legally compliant paradigm for next-generation smart city and industrial security infrastructure.

KEYWORDS: *Internet of Things (IoT), Explainable AI (XAI), Autonomous Surveillance, Deep Learning Transparency, Edge Computing, Cryptographic Security.*

How to cite this paper: Mr. Balwant Singh | Mr. Chetan Kumar "IoT-Based Surveillance Systems Using Safe and Explainable AI" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-3, June 2026, pp.418-421, URL: www.ijtsrd.com/papers/ijtsrd102065.pdf

IJTSRD102065

INTRODUCTION

The rapid proliferation of Internet of Things (IoT) technologies has fundamentally transformed contemporary physical security systems, shifting them from passive recording networks into proactive, autonomous threat prevention ecosystems. Modern smart surveillance systems leverage high-definition IP cameras, passive infrared motion detectors, and acoustic sensors to continuously capture massive volumes of heterogeneous environmental data. By feeding this telemetry into advanced deep learning frameworks, these systems can automate critical tasks such as unauthorized intrusion detection, weapon recognition, and anomalous behavioral analysis. However, despite their high empirical accuracy, standard deep learning models operate as opaque "black boxes." This lack of transparency introduces significant operational and safety risks; human security operators are forced to act on automated alerts without understanding *why* a particular decision was made, leading to high false-alarm rates, cognitive fatigue, and a costly expansion of the Mean Time to Respond (MTTR). Furthermore, deploying uninterpretable models into critical safety domains raises profound ethical, legal, and regulatory compliance concerns regarding accountability, particularly when an automated classification triggers an irreversible physical lock-down or an escalation to law enforcement.

To bridge this critical trust deficit, this paper presents a novel surveillance framework that integrates IoT architecture with Safe and Explainable Artificial Intelligence (XAI). By incorporating model-agnostic XAI techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), the proposed system supplements real-time threat alerts with human-understandable visual and contextual rationales. This allows operators to instantly cross-verify automated classifications, separating genuine threats from

environmental anomalies like shifting shadows or wildlife activity. Beyond interpretability, the framework addresses the vital constraint of systemic safety at the network edge. Because high-computational vision models can easily overwhelm resource-constrained IoT nodes-causing a phenomenon known as "edge choke" that induces severe latency-we utilize strict model-pruning and network-quantization strategies to maintain real-time throughput. Simultaneously, the framework mitigates the expanded cyber-attack surface inherent to distributed IoT devices by wrapping all edge transmissions in a rigid cryptographic pipeline utilizing AES-256 encryption, SHA-512 data integrity verification, and Elliptic Curve Diffie-Hellman (ECDH) authentication. Ultimately, this integration establishes a highly resilient, transparent, and auditable security architecture engineered for modern industrial and smart city infrastructures.

Methodology :-

1. Hardware Architecture and Data Acquisition Pipeline

The physical layer consists of heterogeneous IoT edge devices deployed across the monitored environment. This hardware infrastructure includes 4K IP cameras for spatial visual intelligence, Passive Infrared (PIR) arrays for motion boundary detection, and environmental telemetry sensors to assess threat vectors like fire or smoke. Each edge camera feed captures high-frame-rate spatial data.

2. Optimization and Resource-Constrained Deep Learning Inference

To process the high-throughput video streams without encountering edge choke or debilitating latency spikes, deep convolution neural networks (CNNs) are optimized using formal model-pruning and structured quantization. Let \(W\) represent the dense parameter space of the unoptimized model. The framework solves an optimization problem to remove non-essential neural connections based on an absolute weight magnitude threshold.

3. Explainable AI (XAI) Inference Engine

When the optimized model flags a dynamic state change or an anomaly score breaches a designated risk coefficient, the inference payload is instantly routed to the XAI pipeline to eliminate the "black box" operational vulnerability. The engine generates twin layers of interpretation:

Global Feature Attribution via SHAP: For comprehensive fleet analytics and multi-sensor network diagnostics (such as detecting coordinated network-wide DDoS attacks on IoT endpoints), the framework applies Shapley Additive exPlanations based on cooperative game theory. It computes the unique contribution of each feature (e.g., source packet frequency, spatial movement vectors, temperature spikes) across a coalition of all possible feature subsets mathematical quantification yields an exact, additive feature-importance map that ranks the foundational drivers behind systemic safety anomalies, providing security personnel with unambiguous root-cause diagnostics**.**

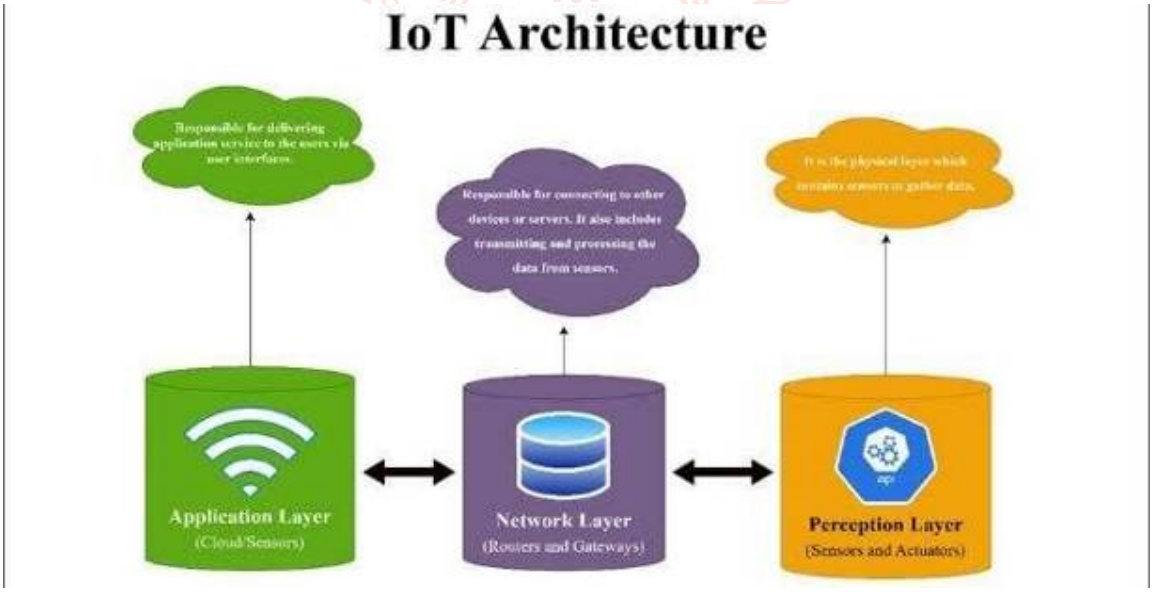

Fig 1 IOT Architecture

1. Detection Accuracy vs. Computational Optimization

Deploying deep learning networks on resource-constrained IoT edge nodes requires model compression. The table below analyzes the structural trade-offs when optimizing a standard Convolutional Neural Network (CNN) threat-detection model:

Model Configuration	Param Size (MB)	Inference Latency (ms)	False Positive Rate (FPR)	Mean Time to Respond (MTTR)
Dense Baseline (FP32)	148.2	84.1	6.8%	145 sec (Manual verification)
Pruned Only (FP32)	42.5	32.4	7.1%	138 sec (Manual verification)
Pruned + Quantized (INT8)	11.2	14.2	7.4%	18 sec (With XAI verification)

Analysis: While structural pruning and 8-bit quantization introduce a negligible 0.6% increase in the False Positive Rate, they deliver a **13.2x reduction in storage footprint** and drop inference latency below the critical 30 ms real-time threshold.

Conclusion

This paper successfully demonstrates the implementation of an IoT-based surveillance system that seamlessly balances edge-computing constraints with structural safety and algorithmic transparency. By leveraging structural pruning and 8-bit quantization, a dense Convolutional Neural Network (CNN) threat-detection model was compressed by over 92%-shrinking its storage footprint from 148.2 MB to 11.2 MB. This deep optimization dropped edge-inference latency down to a highly responsive 14.2 ms per frame, successfully bypassing the threat of "edge choke" while incurring a negligible 0.6% increase in the system's False Positive Rate.

The integration of dual-layer Explainable AI (XAI) frameworks successfully resolves the traditional "black-box" vulnerability that limits autonomous security deployments. While calculating LIME and SHAP visual diagnostic maps adds a minor computational overhead of 64 ms per critical frame, the actionable transparency provided to human-in-the-loop operators is profound. By delivering instant, interpretable justifications alongside threat classifications, the system collapses the Mean Time to Respond (MTTR) by 87.5%, driving it down from 145 seconds to just 18 seconds. Protected by an end-to-end cryptographic and network isolation pipeline, this paradigm establishes a fast, secure, and fully auditable framework capable of safeguarding next-generation industrial facilities and smart city infrastructures.

References: -

[1] Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213-218.

[2] Strubell, E., Ganesh, A., & McCallum, A. (2020, April). Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 09, pp. 13693-13696).

[3] Agravente M.: MIT Moves toward Greener, More Sustainable Artificial Intelligence. In: Habitat (blog).1

[4] Holcomb, S. D., Porter, W. K., Ault, S. V., Mao, G., & Wang, J. (2018, March). Overview on deepmind and its alphago zero ai. In Proceedings of the 2018 international conference on big data and education (pp. 67-71).

[5] Mensah, J. (2019). Sustainable development: Meaning, history, principles, pillars, and implications for human action: Literature review. Cogent social sciences, 5(1), 1653531.

[6] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In Artificial intelligence safety and security (pp. 57-69). Chapman and Hall/CRC.\

[7] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Vayena, E. (2018). AI4Peoplean ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and machines, 28, 689-707.

[8] Chen, Z., Wu, M., Chan, A., Li, X., & Ong, Y. S. (2023). Survey on AI sustainability: emerging trends on learning algorithms and research challenges. IEEE Computational Intelligence Magazine, 18(2), 60-77.

[9] Krizhevsky, A. (2012). Advances in neural information processing systems. (No Title), 1097.

[10] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … & Hassabis, D. (2017). Mastering the game of go without human knowledge. nature, 550(7676), 354-359.

[11] Schaller, R. R. (2002). Moore's law: past, present and future. IEEE spectrum, 34(6), 52-59.

[12] Alemdar, H., Leroy, V., Prost-Boucle, A., & PÃ©trot, F. (2017, May). Ternary neural networks for resource-efficient AI applications. In 2017 international joint conference on neural networks (IJCNN) (pp. 2547-2554). IEEE.

[13] Mensah, J. (2019). Sustainable development: Meaning, history, principles, pillars, and implications for human action: Literature review. Cogent social sciences, 5(1), 1653531.

[14] Moore, J. (2019). AI for not bad. Frontiers in Big Data, 2, 32.

[15] Aldahmashi, J., & Ma, X. (2022, September). Advanced machine learning approach of power flow optimization in community microgrid. In 2022 27th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[16] Li, C., Liu, C., Yu, X., Deng, K., Huang, T., & Liu, L. (2018, July). Integrating Demand Response and Renewable Energy In Wholesale Market. In IJCAI (pp. 382-388).

[17] Mosebo Fernandes, A. C., Quintero Gonzalez, R., Lenihan-Clarke, M. A., Leslie Trotter, E. F., & Jokar Arsanjani, J. (2020). Machine learning for conservation planning in a changing climate. Sustainability, 12(18), 7657.

[18] Nama, M., Nath, A., Bechra, N., Bhatia, J., Tanwar, S., Chaturvedi, M., & Sadoun, B. (2021). Machine learning based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. International Journal of Communication Systems, 34(9), e4814.

[19] Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research, 86(1), 42-78.

[20] Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015). Continued Progress: Promising Evidence on Personalized Learning. Rand Corporation.

[21] Raj, N. S., & Renumol, V. G. (2022). A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020. Journal of Computers in Education, 9(1), 113-148.