IJCRM

International Journal of Contemporary Research In Multidisciplinary

Review Article

Adaptive Edge AI for Context-Aware Inference in Smart Cities

Mr. A. Narayanan [1*], Dr. T. Nagarathinam [2]

[1] Assistant Professor in Computer Science, Swami Dayananda College of Arts & Science, Manjakkudi, Thiruvarur Dt., Tamilnadu. India

[2] Assistant Professor in Computer Science, Swami Dayananda College of Arts & Science, Manjakkudi, Thiruvarur Dt., Tamil Nadu, India

Corresponding Author: *Mr A. Narayanan

DOI: https://doi.org/10.5281/zenodo.17389798

Abstract

As smart cities become increasingly dynamic and data-rich, real-time decision-making at the edge is essential to meet latency, privacy, and bandwidth requirements. Traditional cloud-based AI systems are limited by high communication overhead and a lack of context sensitivity. This paper proposes a novel Adaptive Edge AI framework that performs context-aware inference directly on resource-constrained edge devices. The system dynamically adapts AI model behaviour based on changing urban conditions such as traffic density, pedestrian flow, air quality, or environmental context through a lightweight, modular inference engine. Results show improvements in inference accuracy, latency, and resource efficiency compared to static edge AI deployments. This work offers a scalable approach to deploying intelligent, context-aware services across city infrastructure while maintaining local autonomy and robustness.

Manuscript Information

- **ISSN No:** 2583-7397
- **Received:** 04-08-2025
- **Accepted:** 28-09-2025
- **Published:** 19-10-2025
- **IJCRM:**4(5); 2025: 405-408

- **Plagiarism Checked:** Yes
- **Peer Review Process:** Yes

How to Cite this Article

Narayanan A, Nagarathinam T. Adaptive Edge AI for Context-Aware Inference in Smart Cities. Int J Contemp Res Multidiscip. 2025;4(5):405-408.

Access this Article Online

www.multiarticlesjournal.com

KEYWORDS: Adaptive Edge AI, Smart Cities, Federated Learning, Context-Aware Computing, Dynamic Resource Allocation

1. INTRODUCTION

The evolution of smart cities hinges on the integration of intelligent systems that can sense, analyze, and respond to dynamic urban environments in real time. From traffic management and public safety to environmental monitoring and energy optimization, the demand for intelligent, low-latency Decision-making is increasing rapidly. Traditional AI solutions typically rely on cloud computing infrastructures, which, while Powerful, often suffer from high latency, limited scalability, data privacy concerns, and excessive bandwidth usage. These limitations are especially critical in smart city applications that require real-time responsiveness and context-sensitive operation.

2. LITERATURE REVIEW

The integration of Edge AI in smart city infrastructures has gained significant attention due to the growing need for real-time, low-latency, and privacy-preserving AI applications. This section reviews key areas relevant to this research: Edge AI in smart cities, context-aware computing, adaptive AI models, and existing limitations in current approaches.

2.1. Edge AI in Smart Cities

Edge AI refers to the deployment of AI models on edge devices such as sensors, smartphones, cameras, and embedded systems located closer to the data source. This paradigm reduces communication delays, minimizes reliance on cloud computing, and improves scalability and data privacy. Prior work has demonstrated the utility of edge computing in various smart city domains:

- Traffic surveillance and control systems have used edge-based CNNs to detect congestion and adjust signal timing in real time [Zhao et al., 2022].
- Environmental monitoring systems apply lightweight models to detect air quality anomalies using local sensor data [Lee et al., 2021].

2.2. Context-Aware Computing

Context-aware computing involves systems that can sense and respond to environmental, spatial, or user-related changes. In the smart city context, context may include traffic flow, weather conditions, pollution levels, pedestrian density, or time of day. Several frameworks (e.g., CONTESS [Khezrian et al., 2023]) focus on contextual sensor activation to conserve energy. Others use context data to prioritize resource allocation or data transmission.

2.3 Adaptive AI and Model Updating

The concept of adaptive AI, where models modify their structure or parameters during deployment, is being explored in areas like meta-learning, online learning, and neural architecture adaptation.

- Online model pruning and quantization-aware training have been proposed to reduce model size based on runtime constraints [Han et al., 2020].
- Federated learning allows edge devices to collaboratively train models, but updates are often slow and not context-driven.

3. ARCHITECTURAL DESIGN OF ADAPTIVE EDGE AI SYSTEMS

3.1. Edge-cloud collaboration models

In the evolving landscape of smart cities, vast volumes of heterogeneous data are generated in real-time by distributed sensors, IoT devices, and mobile systems. While cloud computing offers centralized, high-performance processing capabilities, it introduces latency, bandwidth constraints, and privacy concerns that limit its effectiveness for time-critical urban applications. Conversely, edge computing enables localized processing near the data source but suffers from limited computational resources and scalability issues. This dichotomy necessitates a robust edge-cloud collaboration model that dynamically distributes computation, adapts to context, and ensures low-latency, privacy-preserving, and energy-efficient inference.

3.2. Dynamic Resource Allocation and Model Offloading

Smart cities rely on a growing network of connected devices, sensors, and AI-enabled systems to support real-time services such as traffic control, public safety, and environmental monitoring. These applications demand low-latency, high-accuracy inference, often under strict energy and computational constraints. Edge computing brings processing closer to the data source, but edge devices are typically resource-constrained. Conversely, cloud servers offer more power but introduce latency and privacy concerns. This tension creates a critical need for dynamic resource allocation and intelligent model offloading strategies that can adapt to fluctuating network conditions, varying workloads, and evolving context.

3.3. Scalability and Modularity in Edge AI Deployment

As smart cities continue to expand, the number and diversity of deployed edge devices grow exponentially, spanning applications such as traffic analytics, waste management, public safety, and environmental monitoring. These systems require AI capabilities that are not only efficient and real-time but also scalable and modular to accommodate evolving infrastructure, heterogeneous hardware, and diverse data sources. This paper investigates the design and deployment of scalable and modular edge AI frameworks capable of supporting large-scale, heterogeneous smart city environments. The goal is to develop systems where AI components can be independently deployed, updated, and scaled based on demand and context, ensuring seamless integration, interoperability, and efficient resource utilization across dynamic urban infrastructures. T. Nagarathinam (2016), etc., demonstrate A survey on cluster analysis techniques for plant disease diagnosis to detect the diseases in the paddy leaf plant.

4. LEARNING PARADIGMS AND ADAPTATION MECHANISMS

4.1. Federated learning and continual learning at the edge

Smart cities generate massive, continuous streams of sensor and user data distributed across numerous edge devices such as traffic cameras, mobile phones, and environmental sensors. Centralized AI training in such settings is often impractical due to privacy concerns, network limitations, and high data transmission costs. Moreover, urban environments are non-stationery and context-dependent, requiring AI models to adapt continually to evolving patterns like traffic fluctuations, seasonal trends, or emergency events. Federated learning offers a decentralized approach that trains models collaboratively across edge devices without transferring raw data, addressing privacy and bandwidth concerns. However, conventional federated learning methods assume static data distributions and fixed model architectures, which are poorly suited to dynamic city environments. R. P. Ponnusamy et al. (2025) review an article titled as Deep learning with YOLO for smart agriculture: A review of plant leaf disease detection in this survey. They demonstrate how YOLO can be used for agriculture.

4.2. Personalization and domain adaptation strategies

Smart city applications operate in diverse, dynamic, and often highly localized environments ranging from varying traffic patterns across districts to different weather conditions, noise levels, and human behavior across neighborhoods. AI models deployed at the edge must therefore adapt to these unique, location-specific contexts to maintain high accuracy and relevance. However, most pre-trained or centrally trained models lack generalization across domains and fail to account for user-specific or region-specific variations. This paper addresses the need for lightweight, scalable personalization and domain adaptation strategies for edge AI systems in smart cities. The focus is on techniques such as meta-learning, few-shot adaptation, and transfer learning that can dynamically tailor models to individual users, locations. M. Balasubramanian (2018) and others extend a K-NN classifier for plant leaf disease recognition with notable accuracy.

4.3. Lightweight model updates and on-device training

AI-powered smart city systems must continuously adapt to dynamic urban conditions such as shifting traffic patterns, environmental changes, or emerging user behaviors. Traditional AI workflows, which rely on centralized model training and infrequent updates, are too slow and inflexible to support real-time, context-aware services at the edge. At the same time, edge devices, despite being closer to data sources, have limited processing power, memory, and energy capacity, making it impractical to run full-scale training or frequent heavy model updates. This creates a pressing need for lightweight, efficient strategies for model updates and on-device training, enabling real-time adaptation without compromising performance or depleting device resources.

5. CHALLENGES, APPLICATIONS, AND FUTURE DIRECTIONS

5.1. Security, Privacy, and Energy Constraints

The deployment of AI at the edge in smart city environments introduces significant benefits in terms of low-latency decision-making and localized intelligence. However, it also exposes new vulnerabilities and operational challenges. Edge devices operate in open, often untrusted environments, making them susceptible to physical tampering, adversarial attacks, and data breaches. Moreover, smart city applications frequently involve sensitive personal or behavioral data (e.g., facial recognition, location tracking), requiring strong privacy guarantees, especially under regulatory frameworks such as GDPR or CCPA. This paper addresses the fundamental challenge of balancing security, privacy, and energy efficiency in edge AI systems for smart cities. The goal is to design lightweight, secure, and privacy-aware AI frameworks that can operate within tight energy budgets while safeguarding data integrity and user trust.

5.2. Case studies: traffic control, public safety, environmental monitoring

Smart city ecosystems depend on the seamless operation of real-time, intelligent systems across multiple domains, including traffic control, public safety, and environmental monitoring. These domains generate continuous, high-volume data streams from distributed IoT devices such as traffic cameras, acoustic sensors, pollution monitors, and surveillance drones. Centralized cloud-based solutions are inadequate for such time-sensitive tasks due to latency, bandwidth limitations, and privacy concerns. To meet the demand for low-latency, context-aware decision-making, Edge AI has emerged as a promising solution, allowing localized inference and action closer to the data source. However, deploying edge intelligence across diverse domains in a smart city setting poses domain-specific challenges:

- In *traffic control*, AI models must quickly adapt to dynamic vehicle patterns, weather disruptions, and road incidents while minimizing congestion and emissions.
- In *public safety*, edge systems must detect and respond to anomalies or threats in real time, such as identifying gunshots, unusual crowd behavior, or unauthorized access.
- In *environmental monitoring*, edge AI must handle sparse or noisy sensor data to provide accurate, timely updates on air quality, noise pollution, or hazardous conditions.

CONCLUSION

The integration of adaptive Edge AI in smart cities holds transformative potential for enabling real-time, context-aware inference that addresses the growing demands of urban environments. Through dynamic resource allocation, efficient model offloading, and scalable modular architectures, edge intelligence can overcome the inherent limitations of centralized cloud systems, providing low-latency and privacy-preserving solutions. Federated and continual learning approaches further empower edge devices to personalize and adapt AI models locally, ensuring robustness in the face of heterogeneous and evolving urban data. Future research must continue to refine these adaptive frameworks, focusing on resilient architectures, cross-domain generalization, and seamless integration with existing urban infrastructure. Ultimately, the convergence of adaptive edge AI and smart city technologies promises to enhance urban livability, sustainability, and safety, driving the next generation of intelligent cities.

REFERENCES

1. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: Vision and challenges. IEEE Internet Things J. 2016;3(5):637–46.
2. Satyanarayanan M. The emergence of edge computing. Computer. 2017;50(1):30–9.
3. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Trans Intell Syst Technol. 2019;10(2):1–19.

4. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50–60.
5. Deng R, Zhang J, Letaief KB. Edge AI: On-demand accelerating deep neural network inference via edge computing. IEEE Trans Wireless Commun. 2020;19(1):447–57.
6. Wang S, Tuor T, Salonidis T, Leung KK, Makaya C, He T, Chan K. Adaptive federated learning in resource-constrained edge computing systems. IEEE J Sel Areas Commun. 2019;37(6):1205–21.
7. Chen X, Zhang H, Zhang C, Letaief KB. A survey on model compression and acceleration for deep neural networks. IEEE Signal Process Mag. 2020;37(1):46–60.
8. Chen J, Ran X, Zhao J. Deep learning with edge computing: A review. Proc IEEE. 2021;109(8):1341–62.
9. Li X, Huang K, Yang W, Wang S, Zhang Z. On-device federated learning: Communication-efficient and privacy-preserving. IEEE Trans Neural Netw Learn Syst. 2020;31(12):5464–77.
10. Suciu G, Dumitrascu S, Bouillet E. Personalized and adaptive edge AI: Challenges and solutions. IEEE Commun Mag. 2020;58(12):60–5.
11. Ma C, Huang L, Wang J, Yang J. Dynamic resource allocation for edge intelligence: A survey. IEEE Commun Surv Tutorials. 2021;23(2):1113–43.
12. Lin Y, Li X, Yu FR. Privacy preservation for edge computing-enabled smart cities: A survey. IEEE Commun Surv Tutorials. 2019;21(2):1828–60.
13. Aazam M, Huh EN. Fog computing micro data center-based dynamic resource estimation and pricing model for IoT. IEEE Trans Cloud Comput. 2018;7(2):451–64.
14. Liu Y, Kang J, Jin D, Wang J. Context-aware computing in IoT and smart cities: A survey. IEEE Internet Things J. 2021;8(3):2106–31.
15. Xu X, Ma Y, Zhang C. Energy-efficient AI inference in edge computing: A survey. IEEE Access. 2020;8:141360–78.
16. Nagarathinam T, Rameshkumar K. A survey on cluster analysis techniques for plant disease diagnosis. SSRG Int J Comput Sci Eng. 2016;3(6):11–7.
17. Ponnusamy RP, et al. Deep learning with YOLO for smart agriculture: A review of plant leaf disease detection. Int J Comput Tech. 2025;12(4):1–7. Available from: https://ijctjournal.org/
18. Balasubramanian M. K-NN classifier for plant leaf disease recognition. J Appl Sci Comput. 2018;5(11):1–5.

About the Corresponding Author

Mr. A. Narayanan is an Assistant Professor in Computer Science at Swami Dayananda College of Arts & Science, Manjakkudi, Thiruvarur District, Tamil Nadu, India. His academic interests include artificial intelligence, data analytics, and software engineering. He is actively involved in teaching, research, and curriculum innovation, focusing on integrating emerging technologies into higher education to enhance learning outcomes and technical proficiency among students.