

## Adaptive Edge-to-Cloud AI Inference Frameworks for Low-Latency Decision Support

Dr. Mustafa Al-Sadi

Department of Artificial  
Intelligence and Information Systems Iraq

Dr. Rana Al-Jubouri

Department of Computer  
Science and AI Technologies Iraq

VOLUME 03 ISSUE 08 (2026)

Published Date: 20 August 2026 // Page no.: - 85-91

## ABSTRACT

The increasing deployment of artificial intelligence (AI) in healthcare and other time-sensitive decision environments has created a need for inference architectures that simultaneously provide low latency, scalability, privacy, interpretability, and operational resilience. Conventional cloud-centric AI pipelines offer substantial computational capacity but may introduce communication delays, bandwidth dependence, privacy exposure, and reduced responsiveness when large volumes of heterogeneous data must be processed continuously. This research and review paper examines adaptive edge-to-cloud AI inference frameworks as an architectural approach for addressing these limitations. The study synthesizes the provided literature on healthcare big-data management, multimodal biomedical AI, explainable AI, federated learning, privacy-preserving architectures, collaborative neural networks, blockchain-enabled medical analysis, and secure access control. Based on this synthesis, an adaptive framework is conceptually developed around workload-aware inference placement, dynamic model selection, privacy-preserving distributed learning, explainability, and resilient edge-to-cloud orchestration. The analysis indicates that effective edge-to-cloud inference should not be understood merely as moving computation closer to data; rather, it requires coordinated decisions concerning where inference occurs, which model is deployed, what information is transmitted, and how privacy and interpretability requirements influence execution. The proposed framework provides a theoretical foundation for low-latency decision support while recognizing trade-offs involving computational constraints, model complexity, communication overhead, security, and consistency. The findings further indicate that adaptive orchestration is particularly important for healthcare applications in which real-time responsiveness must coexist with stringent privacy and clinical interpretability requirements.

**Keywords:** Edge computing, cloud computing, artificial intelligence, AI inference, low-latency decision support, explainable AI, federated learning, healthcare analytics, privacy preservation, adaptive computing.

## INTRODUCTION

Artificial intelligence has become increasingly important for extracting actionable information from complex and rapidly generated datasets. In healthcare, this development is particularly significant because modern clinical environments combine medical images, physiological signals, electronic records, sensor streams, and other heterogeneous data sources. Big-data systems have therefore become fundamental to healthcare analytics, although their effectiveness depends on the ability to manage large-scale data while preserving security, privacy, and operational reliability (Dash, 2019). The emergence of multimodal biomedical AI further demonstrates that meaningful decision support increasingly requires the integration of multiple data modalities rather than isolated analysis of individual datasets (Acosta, 2022).

A conventional cloud-centric inference architecture typically transfers data from an endpoint or sensing device to a remote computational infrastructure, executes an AI model, and returns the inference result. This approach benefits from centralized processing capacity, but it can become problematic when applications require rapid decisions. Network latency, variable bandwidth, congestion, data-transfer costs, and dependency on continuous connectivity can affect responsiveness. These constraints are particularly relevant to healthcare monitoring systems, where delayed alerts can reduce the practical value of otherwise accurate AI models. Architectures combining Internet-of-Things devices, big-data infrastructures, and healthcare monitoring have consequently emphasized secure and continuous processing capabilities (Manogaran, 2018).

Edge computing addresses part of this challenge by placing computational resources closer to data-generating devices. However, edge devices generally possess less processing

power, memory, and energy capacity than cloud infrastructure. Deploying complex models entirely at the edge may therefore reduce latency while increasing resource pressure. Conversely, placing all inference in the cloud maximizes computational flexibility but can increase communication dependence. The central architectural problem is consequently not whether edge or cloud computing is superior, but how inference workloads can be adaptively distributed across both environments.

This paper investigates that problem through the concept of an adaptive edge-to-cloud AI inference framework. The objective is to develop a research-driven architectural model in which inference placement, model complexity, communication, privacy, and explainability are jointly considered. The analysis is particularly oriented toward intelligent healthcare decision support because the provided literature offers substantial evidence concerning privacy-preserving federated learning, explainable AI, secure healthcare systems, and distributed medical-image analysis.

The conceptual premise is consistent with the broader view that resilient AI inference pipelines must dynamically coordinate computational resources across distributed environments rather than relying on a fixed processing location. Kumar's work on resilient edge-to-cloud AI inference pipelines similarly emphasizes architectural coordination for real-time decision systems, providing an important conceptual basis for treating inference placement as an adaptive systems problem rather than a simple deployment choice (Kumar, 2025).

## LITERATURE REVIEW

### 2.1 Big-Data and Distributed Healthcare Foundations

The foundation for distributed AI inference can be traced to the increasing scale and complexity of healthcare data. Dash identifies healthcare big data as involving significant challenges in data management, analysis, and future scalability (Dash, 2019). This perspective is important because inference architecture is ultimately constrained by the characteristics of the data it processes. A continuously generated stream from connected medical devices, for example, has different computational and communication requirements from a static medical-image dataset.

Manogaran presents an architecture integrating the Internet of Things and big-data infrastructure for secured smart healthcare monitoring and alerting (Manogaran, 2018). The significance of this work for edge-to-cloud inference lies in its implicit architectural relationship between sensing, processing, security, and alert generation. A modern adaptive framework can extend this

principle by assigning immediate processing tasks to edge nodes while retaining cloud resources for computationally demanding or analytically complex operations.

Abouelmehdi et al. emphasize security and privacy challenges associated with big data in healthcare (Abouelmehdi, 2017). These concerns directly affect edge-to-cloud architectures because distributing computation does not automatically eliminate privacy risks. Instead, it creates additional communication and coordination points at which sensitive information may be exposed.

### 2.2 Multimodal and Explainable AI

Acosta demonstrates the growing relevance of multimodal biomedical AI, where multiple forms of clinical information can contribute to AI-supported analysis (Acosta, 2022). Multimodality creates an important inference-placement challenge. Some modalities may be processed locally because they are time-sensitive, while other features can be transferred to centralized infrastructure for deeper analysis. An adaptive architecture must therefore consider not only computational location but also modality-specific processing requirements.

Explainability introduces another architectural dimension. Ali reviews explainable artificial intelligence and emphasizes the broader challenge of achieving trustworthy AI (Ali, 2023). In healthcare, model accuracy alone is insufficient when clinical users need interpretable reasoning. Bharati et al. specifically connect explainable AI with the Internet of Health Things, highlighting the relevance of interpretability within connected healthcare environments (Bharati et al., 2023).

Saraswat et al. further examine explainable AI for Healthcare 5.0 and identify opportunities and challenges associated with intelligent healthcare environments (Saraswat, 2022). These studies collectively indicate that explainability should not be treated as an optional post-processing layer. It can influence which model is deployed, what information is retained, and how results are presented to decision-makers.

### 2.3 Privacy-Preserving Distributed Intelligence

Federated learning provides a mechanism for distributed AI development in which data can remain at participating sites while model information is coordinated across locations. Chen et al. investigate channel-adaptive and sparsity-personalized federated learning for privacy protection in smart healthcare systems (Chen et al., 2024). This work is particularly relevant to adaptive edge-to-cloud architectures because it demonstrates that communication efficiency and privacy can be considered jointly.

Tang et al. address privacy-preserving federated learning with domain adaptation for multi-disease ocular disease recognition (Tang et al., 2023). Their work demonstrates that distributed learning must account for differences between data domains, meaning that a globally optimized model may not

always perform uniformly across edge environments. This supports the need for adaptive model strategies rather than rigid centralized deployment.

Mu et al. combine causal learning and blockchain for explainable federated medical-image analysis (Mu et al., 2024). This study strengthens the theoretical connection between distributed learning, explainability, and trust mechanisms. However, additional security layers may introduce computational and communication overhead, creating a trade-off that an adaptive inference framework must explicitly manage.

## 2.4 Secure and Collaborative AI Inference

Ahmad et al. present a secure and interpretable AI system for epilepsy diagnosis using EEG signals (Ahmad et al., 2024). The work illustrates the need to combine security and interpretability in clinical AI rather than optimizing inference accuracy independently of deployment requirements.

Khan et al. propose a collaborative SRU network incorporating dynamic behavior aggregation, reduced communication overhead, and explainable features (Khan et al., 2024). The combination is highly relevant to edge-to-cloud environments because recurrent and collaborative architectures can potentially support sequential data processing while limiting communication demands.

Devi et al. focus on explainable reverse verification of MRI-image classification by clinical experts (Devi et al., 2023). Their work highlights an important distinction between computational inference and clinically meaningful validation. An adaptive framework must therefore support not only fast prediction but also mechanisms through which predictions can be examined or verified.

Finally, Liang et al. investigate privacy-preserving authentication and flexible revocation for smart healthcare access control (Liang et al., 2024). Their findings reinforce the requirement for identity and authorization mechanisms in distributed healthcare architectures. Edge-to-cloud inference cannot be considered secure merely because data processing is decentralized; access control must operate across the distributed computational hierarchy.

## 2.5 Research Gap and Theoretical Positioning

The provided literature strongly addresses individual dimensions of intelligent healthcare systems, including big-data processing, explainability, federated learning, security, privacy, and collaborative AI. However, these dimensions are frequently studied as distinct technical problems. The principal research gap is therefore the need for an integrated inference framework that treats latency,

resource availability, privacy, explainability, communication cost, and model selection as interdependent variables.

The theoretical position of this paper is that adaptive edge-to-cloud inference should be modeled as a multi-objective orchestration problem. The optimal execution location is not fixed; it depends on the characteristics of the task, available resources, network conditions, data sensitivity, model complexity, and required response time. This conceptualization aligns with resilient edge-to-cloud inference principles in which distributed resources are coordinated according to real-time decision requirements (Kumar, 2025).

## METHODOLOGY

### 3.1 Research Design

This study adopts a structured research-and-review methodology based exclusively on the provided literature. Rather than conducting an empirical benchmark, the research synthesizes architectural principles from the selected studies and translates them into a conceptual adaptive edge-to-cloud inference framework.

The methodology consists of four analytical stages: identification of system requirements, decomposition of inference functions, adaptive orchestration design, and evaluation through qualitative performance dimensions. The approach is appropriate for an architectural study because the primary objective is to establish a coherent framework connecting distributed inference with privacy, explainability, and resilience.

### 3.2 Adaptive Edge-to-Cloud Architecture

The proposed architecture consists of five conceptual layers: data acquisition, edge inference, adaptive orchestration, cloud intelligence, and decision support.

The data-acquisition layer receives information from sensors, medical devices, imaging systems, and other endpoints. Initial processing occurs near the source to remove unnecessary transmission and enable rapid response. This principle follows the architectural logic of IoT-enabled healthcare monitoring in which data generation and intelligent alerting must be coordinated (Manogaran, 2018).

The edge-inference layer executes models suitable for local resource constraints. Lightweight classification, anomaly detection, filtering, and preliminary feature extraction can be prioritized at this level. For example, an EEG monitoring device could perform initial signal analysis locally and transmit only relevant features or abnormal segments rather than continuously forwarding the entire raw stream.

The adaptive-orchestration layer functions as the decision mechanism of the framework. It evaluates workload complexity, edge resource utilization, network state, privacy

requirements, and latency constraints. The inference request can then be assigned to an edge device, intermediate node, or cloud infrastructure.

The cloud-intelligence layer provides computationally intensive processing, global model management, large-scale analytics, and cross-device learning. Its role is not eliminated by edge computing; instead, the cloud becomes a higher-level analytical resource that complements local inference.

The final decision-support layer converts model outputs into actionable information. In healthcare, this layer should include interpretable explanations because explainability contributes to trust and clinical evaluation (Ali, 2023; Bharati et al., 2023).

### 3.3 Adaptive Inference Placement

Inference placement can be represented conceptually as a utility optimization problem. For an inference task (i), the orchestration mechanism seeks a deployment location (l) that minimizes a composite cost:

$$C(i,l) = w_L L(i,l) + w_E E(i,l) + w_N N(i,l) + w_P P(i,l) + w_X X(i,l)$$

where (L) represents latency, (E) represents computational or energy cost, (N) represents communication overhead, (P) represents privacy exposure, and (X) represents explainability or interpretability cost. The coefficients ( $w_L, w_E, w_N, w_P, w_X$ ) represent application-specific priorities.

This formulation illustrates why a single universal placement policy is insufficient. A critical emergency alert may assign a very high weight to latency, favoring local inference. A computationally intensive multimodal diagnostic task may place greater emphasis on processing capacity and therefore use cloud resources. A privacy-sensitive workload may favor local or federated processing.

### 3.4 Dynamic Model Selection

Adaptive inference also requires model-level flexibility. A framework may maintain multiple model variants, ranging from lightweight edge models to more computationally intensive cloud models. The lightweight model provides rapid preliminary inference, while the cloud model can perform deeper analysis when sufficient time and connectivity are available.

The collaborative neural-network approach discussed by Khan et al. illustrates how model architecture itself can incorporate communication-efficiency considerations (Khan et al., 2024). Similarly, multimodal AI demonstrates that model selection must consider the availability and relevance of multiple information sources (Acosta, 2022).

A practical example involves medical imaging. An edge node could perform rapid preliminary screening to identify images requiring urgent attention. Complex cases could subsequently be transmitted to a cloud-based model for more detailed analysis. This hierarchical strategy reduces unnecessary network traffic while preserving access to higher computational capacity.

### 3.5 Privacy-Preserving Processing

Privacy should be incorporated into orchestration rather than added after deployment. Sensitive raw data can remain near its source, while only derived representations or model updates are transmitted when appropriate. Federated learning provides a theoretical mechanism for such distributed learning, although privacy preservation does not automatically eliminate all security risks (Chen et al., 2024; Tang et al., 2023).

The proposed framework therefore distinguishes between raw-data transmission, feature transmission, inference-result transmission, and model-update transmission. Each category can receive a different security policy. Authentication and flexible authorization are also necessary because distributed infrastructure creates multiple access points (Liang et al., 2024).

Blockchain-based mechanisms may additionally support trust and traceability in distributed medical analysis, as demonstrated by Mu et al. (2024). Nevertheless, blockchain introduces additional processing and coordination requirements. Consequently, the framework should deploy such mechanisms selectively rather than treating them as universally necessary.

### 3.6 Explainability-Aware Inference

Explainability is incorporated as a first-class architectural requirement. The orchestration layer can select not only a model based on latency and computational capacity but also according to the explanation requirements of the application.

For example, an edge classifier may produce a rapid anomaly score, while a cloud-based model may provide a more detailed explanation for a clinician. This distinction is consistent with research emphasizing trustworthy and interpretable AI in healthcare (Ali, 2023; Saraswat, 2022).

Clinical verification provides another dimension. Devi et al. demonstrate the importance of expert-oriented evaluation of AI classification results in medical imaging (Devi et al., 2023). Therefore, an adaptive system should retain sufficient

evidence to allow critical predictions to be reviewed rather than optimizing exclusively for minimum inference time.

### 3.7 Resilience and Failure Handling

A resilient edge-to-cloud architecture should continue providing degraded but useful functionality when network connectivity or cloud resources become unavailable. Local inference can serve as a fallback mechanism, while queued data can be synchronized after connectivity is restored. This architectural principle is consistent with the need for resilient real-time inference pipelines in distributed decision systems (Kumar, 2025).

However, resilience introduces consistency challenges. Edge models may become outdated, local nodes may have different computational capabilities, and federated updates may not arrive simultaneously. Tang et al.'s consideration of domain adaptation illustrates why distributed model performance can vary across environments (Tang et al., 2023).

## RESULTS

The synthesis identifies five principal findings concerning adaptive edge-to-cloud AI inference.

First, latency is primarily an architectural variable rather than solely a model-performance variable. A highly accurate model may have limited operational value if communication and processing delays prevent timely decisions. Edge inference can therefore provide significant functional benefits for time-sensitive applications by reducing dependence on remote computation.

Second, adaptive placement is more appropriate than fixed edge or cloud deployment. Edge resources are constrained, whereas cloud resources are comparatively scalable but communication-dependent. The most effective architecture therefore dynamically assigns tasks according to workload characteristics, resource conditions, and response requirements. This finding is consistent with resilient edge-to-cloud inference principles emphasizing coordinated distributed processing (Kumar, 2025).

Third, privacy and communication efficiency are closely interconnected. Federated learning and privacy-preserving approaches demonstrate that reducing the movement of raw healthcare data can simultaneously address privacy concerns and communication burdens (Chen et al., 2024; Tang et al., 2023). Nevertheless, distributed learning introduces model-coordination requirements and does not eliminate security risks.

Fourth, explainability must be incorporated into inference architecture. The literature consistently associates

explainability with trustworthy healthcare AI (Ali, 2023; Bharati et al., 2023). An adaptive framework should therefore select models partly according to the level of interpretability required for the decision. The result is a shift from accuracy-centric deployment toward multi-dimensional decision support.

Fifth, security and resilience create necessary trade-offs. Secure authentication, privacy-preserving computation, blockchain, and distributed access controls can improve trust but may increase computational or communication overhead (Abouelmehdi et al., 2017; Liang et al., 2024; Mu et al., 2024). The findings therefore support a policy-driven architecture in which security mechanisms are selected according to risk rather than uniformly applied at maximum intensity.

Overall, the synthesized framework indicates that low-latency AI should be achieved through coordinated optimization of placement, model complexity, communication, privacy, explainability, and resilience rather than through edge deployment alone.

## DISCUSSION

The findings position adaptive edge-to-cloud inference as a multi-objective architectural problem. Existing literature provides strong evidence for each individual component, but the principal contribution of the proposed framework is their integration into a single decision model. This integration is important because optimizing one dimension can negatively affect another. Reducing latency through local computation, for example, may increase energy consumption and reduce available memory, while maximizing model complexity may improve predictive performance but increase response time.

Theoretical implications emerge from the shift toward workload-aware orchestration. Traditional AI deployment can be viewed as a relatively static mapping between a model and computational infrastructure. The adaptive model instead treats infrastructure as a dynamic resource pool. Inference becomes a runtime decision influenced by network conditions, resource availability, data sensitivity, and decision urgency. This perspective is consistent with resilient edge-to-cloud architectural thinking (Kumar, 2025).

The literature also reveals a significant relationship between explainability and system design. Explainability research emphasizes trustworthiness (Ali, 2023), while healthcare-oriented studies connect interpretable models with connected health systems and clinical evaluation (Bharati et al., 2023; Devi et al., 2023). Consequently, the architecture should distinguish between rapid automated screening and high-assurance clinical decision support. The former can prioritize speed, whereas the latter may justify additional computational expenditure to generate interpretable evidence.

Privacy introduces a similar trade-off. Federated learning reduces the necessity of transferring raw datasets, but

distributed learning requires communication among participants and may encounter heterogeneous data distributions (Chen et al., 2024; Tang et al., 2023). Domain adaptation is therefore particularly relevant when edge nodes operate under different patient populations or sensing conditions.

The security literature further suggests that decentralization changes rather than removes security requirements. Big-data environments expose sensitive information to multiple processing stages (Abouelmehdi et al., 2017), while authentication and revocation become increasingly important when healthcare services are distributed across devices and computational nodes (Liang et al., 2024). Blockchain can support traceability and trust, but its additional overhead means that its deployment should be risk-sensitive (Mu et al., 2024).

Several limitations remain. The proposed framework is conceptual and has not been evaluated through controlled latency experiments, energy measurements, or clinical trials. The utility formulation also requires empirical calibration of its weighting parameters. Furthermore, heterogeneous hardware and network conditions may make globally consistent orchestration difficult. Finally, explainability itself can impose computational overhead and may vary substantially across model classes.

These limitations identify an important research direction: future implementations should benchmark adaptive placement policies against edge-only and cloud-only baselines using latency, accuracy, communication volume, energy consumption, privacy risk, and explanation quality as simultaneous evaluation criteria.

## CONCLUSION

This paper examined adaptive edge-to-cloud AI inference frameworks for low-latency decision support through a structured synthesis of the provided literature. The analysis demonstrates that the fundamental challenge is not simply deciding between edge and cloud computing. Instead, effective AI inference requires continuous coordination of computational placement, model complexity, communication, privacy, security, explainability, and resilience.

The proposed framework addresses this challenge through a layered architecture combining local preprocessing, adaptive edge inference, workload-aware orchestration, cloud-based intelligence, and explainable decision support. Its principal theoretical contribution is the conceptualization of inference placement as a multi-objective optimization problem. This approach recognizes that different tasks require different balances between latency, resource utilization, privacy, communication efficiency, and interpretability.

The literature on federated learning, explainable AI, secure healthcare systems, collaborative neural architectures, and distributed medical analysis supports the feasibility of this integrated perspective. At the same time, the synthesis shows that every architectural improvement involves trade-offs. Greater privacy can increase coordination complexity, stronger explainability can increase computational requirements, and greater resilience can require model redundancy and synchronization mechanisms.

For practical deployment, future systems should therefore employ adaptive policies rather than fixed infrastructure assignments. Particularly in healthcare, rapid edge-based screening should be complemented by cloud-supported analytical depth when required. Privacy-preserving learning, secure access control, and explainability should be incorporated at architectural design time rather than treated as secondary additions.

Future research should validate the proposed framework using real-world heterogeneous edge devices and cloud infrastructures. Experimental evaluation should measure latency, model accuracy, energy consumption, communication overhead, privacy preservation, explanation quality, and resilience under network failures. Such evaluation would establish whether adaptive orchestration can consistently achieve the low-latency and trustworthy decision support required by next-generation intelligent healthcare environments.

## REFERENCES

1. K. Abouelmehdi, "Big data security and privacy in healthcare: A review," *Procedia Comput. Sci.*, vol. 113, pp. 73–80, 2017.
2. J. N. Acosta, "Multimodal biomedical AI," *Nature Med.*, vol. 28, no. 9, pp. 1773–1784, 2022.
3. I. Ahmad, M. Zhu, G. Li, D. Javeed, P. Kumar, and S. Chen, "A secure and interpretable AI for smart healthcare system: A case study on epilepsy diagnosis using EEG signals," *IEEE J. Biomed. Health Inform.*, early access, Mar. 20, 2024, doi: 10.1109/JBHI.2024.3366341.
4. S. Ali, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, 2023, Art. no. 101805.
5. S. Bharati, M. R. H. Mondal, P. Podder, and U. Kose, "Explainable artificial intelligence (XAI) with IoHT for smart healthcare: A review," in *Interpretable Cognitive Internet of Things for Healthcare*. Berlin, Germany : Springer, 2023, pp. 1–24.
6. Z. Chen, J. Du, X. Hou, K. Yu, J. Wang, and Z. Han, "Channel adaptive and sparsity personalized federated learning for privacy protection in smart healthcare

systems," IEEE J. Biomed. Health Inform., early access, Jan. 15, 2024, doi: 10.1109/JBHI.2024.3353791.

7. S. Dash, "Big data in healthcare: Management, analysis and future prospects," J. Big Data, vol. 6, no. 1, pp. 1–25, 2019.
8. S. Devi, M. N. Sahoo, and S. Bakshi, "Explainable reverse verification of goodness of classification of MRI images by clinical experts," IEEE J. Biomed. Health Inform., early access, May 26, 2023, doi: 10.1109/JBHI.2023.3280184.
9. I. A. Khan, I. Razzak, D. Pi, U. Zia, S. Kamal, and Y. Hussain, "A novel collaborative SRU network with dynamic behaviour aggregation, reduced communication overhead and explainable features," IEEE J. Biomed. Health Inform., early access, Jan. 10, 2024, doi: 10.1109/JBHI.2024.3352013.
10. X. Liang, Y. Liu, and J. Ning, "An access control scheme with privacy-preserving authentication and flexible revocation for smart healthcare," IEEE J. Biomed. Health Inform., early access, Apr. 24, 2024, doi: 10.1109/JBHI.2024.3391218.
11. G. Manogaran, "A new architecture of internet of Things and Big Data ecosystem for secured smart healthcare monitoring and alerting system," Future Gener. Comput. Syst., vol. 82 : pp. 375–387, 2018.
12. J. Mu, M. Kadoch, T. Yuan, W. Lv, Q. Liu, and B. Li, "Explainable federated medical image analysis through causal learning and blockchain," IEEE J. Biomed. Health Inform., early access, Mar. 18, 2024, doi: 10.1109/JBHI.2024.3375894.
13. D. Saraswat, "Explainable AI for healthcare 5.0: Opportunities and challenges," IEEE Access, vol. 10, pp. 84486–84517, 2022.
14. Z. Tang, H. S. Wong, and Z. Yu, "Privacy-preserving federated learning with domain adaptation for multi-disease ocular disease recognition," IEEE J. Biomed. Health Inform., early access, Aug. 17, 2023, doi: 10.1109/JBHI.2023.3305685.
15. K. Kumar, "Edge-to-Cloud AI Inference Pipelines: A Resilient Architecture for Real-Time Decision Systems," 2025 International Conference on Computer and Applications (ICCA), Bahrain, Bahrain, 2025, pp. 1-6, doi: 10.1109/ICCA66035.2025.11430905.