International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

Journal Website: http://sciencebring.com/index.php/ijasr

Research Article

Resilient Edge-to-Cloud AI Architectures for Distributed Real-Time Decision Making

Submission Date: June 28, 2026, **Accepted Date:** July 22, 2026,
Published Date: August 17, 2026
Crossref doi: https://doi.org/10.37547/ijasr-06-08-07

Dr. Chinedu Eze
Department of Artificial Intelligence Vietnam Institute of Advanced Computing Hanoi, Vietnam

Dr. Fatima Bello
Department of Computer Science and Intelligent Systems Nigeria

Abstract

The increasing deployment of artificial intelligence (AI) in distributed environments has created a need for architectures capable of combining low-latency inference, computational scalability, data protection, and operational resilience. Conventional cloud-centric AI pipelines can provide substantial computational resources but may introduce latency, network dependency, privacy concerns, and vulnerability to service interruptions. Edge-to-cloud architectures address these limitations by distributing data processing and inference across edge devices, intermediate computing nodes, and centralized cloud infrastructure. This paper develops a research-driven conceptual framework for resilient edge-to-cloud AI architectures supporting distributed real-time decision making. The study synthesizes the supplied literature concerning AI-assisted medical image classification, radiological decision processes, ground-truth uncertainty, workload-related behavior, fatigue, information protection, and automated image segmentation. Particular emphasis is placed on how inference placement, data integrity, uncertainty management, workload awareness, and adaptive orchestration can collectively improve system resilience. The methodology develops a layered architectural model consisting of sensing and acquisition, edge inference, adaptive orchestration, cloud intelligence, resilience management, and decision feedback. The analysis indicates that resilience should not be treated exclusively as infrastructure availability; rather, it must encompass model reliability, data quality, human interaction, computational continuity, and decision confidence. The proposed framework provides a conceptual basis for designing distributed AI systems in which latency-sensitive decisions are executed near data sources while computationally intensive and

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

globally coordinated processes remain cloud-enabled. The study also identifies limitations associated with heterogeneous devices, uncertain ground truth, model drift, communication failures, and the absence of uniform evaluation criteria.

KEYWORDS

Edge AI, Cloud Computing, Distributed AI, Real-Time Decision Making, Resilient Architectures, AI Inference, Edge-to-Cloud Computing, Decision Support, Model Reliability

INTRODUCTION

1.1 Background

Artificial intelligence is increasingly incorporated into decision-support environments in which the value of an inference depends not only on predictive accuracy but also on the speed, reliability, and contextual validity of the resulting decision. In medical imaging, for example, automated classification, segmentation, and decision-support systems can assist professionals in interpreting complex data. Research on brain MR-image classification demonstrates the use of deep transfer learning and automated image-processing techniques, while automated MRI segmentation has been investigated for determining brain-tumor target volumes in radiation treatment planning (Behera et al., 2022; Mazzara et al., 2004). These applications illustrate a broader technological requirement: AI systems must operate as dependable computational pipelines rather than as isolated predictive models.

A centralized cloud architecture offers advantages in storage, model management, large-scale computation, and coordinated analytics. However, real-time distributed applications cannot always depend on continuous communication with remote cloud infrastructure. Network latency, bandwidth limitations, intermittent connectivity, and data-transfer requirements can interfere with time-sensitive inference. An edge-to-cloud architecture therefore distributes computational responsibilities between geographically or operationally distributed edge resources and centralized cloud platforms. The architectural objective is not simply to move computation closer to users but to determine which processing tasks should occur locally, which should be delegated, and how the system should recover when individual components become unavailable.

Recent work specifically addressing resilient edge-to-cloud AI inference pipelines emphasizes architectural mechanisms for maintaining real-time decision capability under distributed operating conditions (Kumar, 2025). The significance of resilience becomes clearer when AI is deployed in domains where incorrect or delayed decisions have operational

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

consequences. Studies of radiological workload and gaze behavior show that human performance is itself affected by workload conditions, while research on fatigue demonstrates that radiology involves substantial cognitive and operational pressures (Pershin et al., 2022; Stec et al., 2018). Consequently, an effective AI architecture should consider not only computational performance but also the relationship between machine inference, uncertainty, human interpretation, and decision context.

The central problem addressed in this paper is therefore the design of a resilient architecture capable of sustaining distributed real-time AI inference while balancing latency, computational capacity, data integrity, uncertainty, privacy, and decision reliability. The objective is to synthesize the supplied literature into a conceptual edge-to-cloud framework and identify architectural mechanisms that can support dependable decision making. The scope is primarily conceptual and is informed by evidence from AI-enabled medical and imaging applications, with the proposed architecture designed to be generalizable to other distributed real-time AI environments.

LITERATURE REVIEW

The supplied literature provides complementary evidence for understanding the requirements of resilient AI systems. Behera et al. (2022) demonstrate that brain MR-image classification can benefit from superpixel-based deep transfer learning, highlighting the computational value of advanced feature representation and transfer learning. Devi et al. (2023) further show that situational and instrumental distortions can influence the classification of brain MR images. Together, these studies indicate that AI reliability is influenced by both model architecture and the quality of the conditions under which data are acquired.

The importance of ground truth is another central theme. Moore and Toumpanakis (2021) define ground truth in a radiological context, while Bowler et al. (2020) explicitly examine the implications of ground-truth uncertainty for deep-learning results. This distinction is important for edge-to-cloud systems because distributing computation does not eliminate uncertainty in the underlying training or validation data. A highly available architecture can continue producing predictions even when its reference labels are uncertain. Resilience must therefore incorporate epistemic and data-related uncertainty rather than being reduced to hardware or network availability.

Research on medical imaging also demonstrates the need for robust information processing. Azemin et al. (2020) investigate deep learning for COVID-19 prediction using radiologist-adjudicated chest X-ray data, illustrating how AI models depend on curated and professionally evaluated datasets. Devi et al. (2019) examine methods for hiding medical information in brain MR images without affecting classification accuracy, demonstrating the tension between information protection and computational utility. In an edge-to-cloud architecture, this tension

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

becomes particularly important because distributed data movement can increase exposure while local processing can reduce unnecessary transmission.

The human decision environment is equally relevant. Chang et al. (2020) analyze radiologists' behavior in diagnosing thyroid nodules using data-driven methods, while Pershin et al. (2022) examine workload-related changes in radiologists' gaze patterns. Stec et al. (2018) identify fatigue as a significant issue in radiology, and Kim and Mansfield (2014) discuss delayed diagnoses and perpetuated errors. These findings imply that AI architecture should support human decision makers rather than merely optimize model throughput. An architecture that generates technically accurate predictions but overwhelms users with poorly prioritized alerts may fail operationally.

Mazzara et al. (2004) demonstrate the value of automated MRI segmentation for treatment planning, providing an example in which computational processing directly contributes to a downstream clinical decision. Such applications require reliable data pipelines because segmentation errors can propagate into subsequent planning processes. The literature therefore supports a theoretical position in which resilience emerges from the interaction of computational continuity, data quality, model reliability, uncertainty management, and human-centered decision support.

The major research gap is that the supplied studies largely investigate individual components—classification, segmentation, ground truth, privacy, workload, fatigue, or diagnostic behavior—rather than integrating these dimensions into a single resilient distributed architecture. Kumar (2025) provides a direct architectural perspective on resilient edge-to-cloud AI inference, creating a foundation for integrating these separate requirements. The present study extends this perspective conceptually by treating resilience as a multidimensional property of the entire inference-to-decision pipeline.

METHODOLOGY

3.1 Research Design

This study adopts a conceptual research and structured literature-synthesis methodology. Rather than conducting a new empirical experiment, the methodology extracts architectural requirements from the supplied references and maps them onto an edge-to-cloud AI decision pipeline. The approach is appropriate because the objective is to establish a framework connecting previously studied technical and human factors rather than to claim experimental performance improvements.

The methodological process consists of four analytical stages: identification of system requirements, classification of resilience dimensions, architectural decomposition, and synthesis of decision-flow mechanisms. The first stage identifies requirements related to latency, accuracy, uncertainty, privacy, workload, computational continuity, and human decision

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

support. The second classifies these requirements into data, model, infrastructure, communication, and human-centered resilience. The third maps these dimensions to architectural layers. The final stage evaluates the resulting framework through qualitative comparison with the supplied literature.

3.2 Proposed Edge-to-Cloud Architecture

The proposed architecture is organized into six functional layers: data acquisition, edge intelligence, adaptive orchestration, cloud intelligence, resilience management, and decision feedback.

The data acquisition layer collects information from sensors, imaging devices, monitoring systems, or other distributed sources. Data validation occurs at this stage to identify incomplete, corrupted, duplicated, or anomalous inputs. This requirement follows directly from evidence that distortions can affect AI classification outcomes (Devi et al., 2023). The objective is to prevent unreliable inputs from unnecessarily propagating through the inference pipeline.

The edge intelligence layer performs latency-sensitive preprocessing and inference close to the data source. Lightweight or optimized models can provide immediate predictions without waiting for cloud communication. In an imaging scenario, an edge node could perform image quality assessment and preliminary classification before forwarding selected cases to a cloud-based model. Transfer learning approaches demonstrate how specialized classification can be developed from established model representations (Behera et al., 2022).

The adaptive orchestration layer determines whether a task should remain at the edge, be transmitted to an intermediate node, or be processed in the cloud. Its decision criteria include latency requirements, computational load, connectivity, privacy constraints, prediction confidence, and task complexity. This layer is central to resilience because it prevents dependence on a single computational location. The principle is consistent with the resilient inference perspective described by Kumar (2025).

The cloud intelligence layer provides computationally intensive operations such as large-model inference, global model updates, historical analytics, centralized training, and cross-site aggregation. Cloud processing is particularly appropriate when computational requirements exceed edge-device capacity. However, the architecture should avoid making cloud availability a prerequisite for every decision.

The resilience management layer continuously monitors system health, inference confidence, communication status, resource utilization, and data-quality indicators. If connectivity deteriorates, the system can switch to local inference. If an edge model exhibits low confidence, the case can be escalated to a more computationally capable cloud model. If the cloud becomes unavailable, previously validated local

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

models can provide degraded but operational service.

The decision feedback layer presents predictions, confidence indicators, and relevant supporting information to human decision makers. This layer is essential because research on workload, gaze behavior, fatigue, and diagnostic errors demonstrates that the human environment can influence decision performance (Pershin et al., 2022; Stec et al., 2018; Kim and Mansfield, 2014). Accordingly, resilience should include mechanisms for prioritizing high-risk cases and reducing unnecessary cognitive burden.

3.3 Resilience Model

The proposed framework defines resilience through five interacting dimensions: computational resilience, communication resilience, data resilience, model resilience, and human-centered resilience.

Computational resilience concerns continued operation despite resource exhaustion or node failure. Communication resilience concerns maintaining service during intermittent or degraded network conditions. Data resilience concerns preserving integrity and handling uncertainty. Model resilience concerns maintaining dependable predictions under distributional changes and imperfect inputs. Human-centered resilience concerns delivering information in a manner that supports rather than disrupts professional judgment.

Ground-truth uncertainty is especially important to model resilience. Bowler et al. (2020) demonstrate that uncertainty in reference information can influence the interpretation of deep-learning results. Therefore, a resilient system should distinguish between high-confidence predictions and cases requiring escalation rather than treating every inference as equally reliable.

3.4 Illustrative Decision Flow

Consider a distributed medical-imaging environment in which an imaging device produces a brain MR image. The edge node first performs data-quality validation and preliminary inference. If the image is suitable and the prediction confidence exceeds a predefined threshold, the result can be immediately presented to the decision-support interface. If the model detects distortion or uncertainty, the image can be routed to the cloud for more advanced processing. Automated segmentation may subsequently support treatment-planning workflows, consistent with the role of MRI segmentation demonstrated by Mazzara et al. (2004).

If the network becomes unavailable, the edge node continues operating using its locally deployed model. When connectivity returns, relevant data and inference records can be synchronized with the cloud. This approach creates graceful degradation rather than complete service interruption.

RESULTS

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

The conceptual synthesis produces five principal findings. First, resilience is multidimensional. The reviewed literature indicates that dependable AI cannot be achieved solely through infrastructure redundancy. Classification performance can be affected by image distortions, while ground-truth uncertainty can influence interpretation of model performance (Devi et al., 2023; Bowler et al., 2020). Consequently, infrastructure availability represents only one component of resilience.

Second, edge processing is most valuable for latency-sensitive and locally actionable tasks, whereas cloud resources remain appropriate for computationally intensive or globally coordinated operations. This division reduces dependence on remote infrastructure while retaining access to large-scale computational capacity. The architectural principle aligns with the resilient edge-to-cloud inference approach proposed by Kumar (2025).

Third, uncertainty-aware orchestration is more appropriate than static workload placement. A fixed architecture that permanently assigns all inference to the edge or cloud cannot adapt effectively to changing resource availability, connectivity, or prediction confidence. A dynamic system can retain simple and high-confidence decisions locally while escalating uncertain or computationally demanding cases to centralized infrastructure.

Fourth, human-centered resilience is essential for decision-support applications. Evidence concerning radiologists' workload, gaze behavior, fatigue, and delayed diagnoses suggests that decision quality depends partly on human operating conditions (Pershin et al., 2022; Stec et al., 2018; Kim and Mansfield, 2014). Therefore, the proposed architecture should optimize not merely inference latency but also information prioritization and cognitive usability.

Fifth, privacy and information protection should influence architectural placement. Local processing can reduce unnecessary movement of sensitive information, while controlled cloud synchronization can support centralized analytics. The relevance of this principle is supported by research examining the protection of medical information without compromising classification accuracy (Devi et al., 2019).

Collectively, the findings suggest that resilient edge-to-cloud AI should be designed as an adaptive pipeline rather than as a static collection of edge and cloud resources. The strongest architectural configuration is one in which inference placement, confidence assessment, data validation, and failure recovery operate together. This approach also provides a mechanism for graceful degradation: when the optimal computational path is unavailable, a lower-complexity alternative can preserve essential functionality.

However, the findings are conceptual rather than experimentally validated. No quantitative claims regarding latency reduction, accuracy improvement, energy savings, or failure-recovery time can therefore be made from the supplied evidence. The framework should instead be regarded as a research model that can guide

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

subsequent implementation and empirical evaluation.

Discussion

The proposed framework changes the conventional interpretation of AI resilience from a narrow infrastructure concept to a system-level property. Traditional reliability strategies often emphasize redundant servers, backup storage, or network failover. Such mechanisms remain necessary, but they do not address failures arising from inaccurate inputs, uncertain labels, model degradation, or excessive human workload. The literature demonstrates that these factors can independently influence AI-assisted decision processes, making a broader resilience definition necessary.

One important theoretical implication is that uncertainty should become an architectural control variable. Ground-truth uncertainty can affect the interpretation of AI outcomes (Bowler et al., 2020), while image distortions can affect classification performance (Devi et al., 2023). Instead of treating uncertainty as an output that is displayed after inference, the proposed architecture incorporates it into orchestration. A low-confidence result can trigger cloud escalation, additional processing, or human review. This creates a feedback relationship between model confidence and computational allocation.

A second implication concerns the relationship between automation and human expertise. Studies of radiologists' behavior and workload indicate that professional decision making is influenced by cognitive and operational factors (Chang et al., 2020; Pershin et al., 2022). AI therefore should not simply maximize the number of automated predictions. A resilient decision-support system should identify which outputs are most consequential and allocate attention accordingly. This is particularly important in environments affected by fatigue, where excessive information can reduce rather than increase practical value (Stec et al., 2018).

A third issue concerns the trade-off between local processing and centralized intelligence. Edge inference reduces communication dependence and can support rapid responses, but edge devices typically have constrained computational resources. Cloud systems provide greater computational capacity but introduce communication dependency and potentially greater data exposure. Kumar (2025) frames the edge-to-cloud model as a means of balancing these competing requirements. The present framework extends that balance by incorporating confidence, privacy, and human workload into placement decisions.

There are also significant limitations. Heterogeneous edge devices may support different model architectures and computational capabilities, complicating deployment and synchronization. Model updates may introduce inconsistencies between edge and cloud versions. Privacy-preserving mechanisms may increase computational overhead. Furthermore, resilient fallback models may produce lower-quality predictions than their cloud counterparts. The

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

system must therefore distinguish between maintaining service availability and maintaining equivalent decision quality.

The clinical literature also suggests that automated systems should not be treated as infallible substitutes for professional judgment. Research concerning delayed diagnoses and perpetuated errors demonstrates that errors can persist within diagnostic processes (Kim and Mansfield, 2014). A distributed architecture could potentially amplify such errors if faulty models are replicated across multiple edge nodes. Model governance, version control, validation, and escalation mechanisms are therefore necessary complements to technical resilience.

Finally, future empirical evaluation should examine latency, inference accuracy, recovery time, bandwidth consumption, energy use, model consistency, privacy exposure, and human workload simultaneously. Such multidimensional evaluation would provide stronger evidence for whether edge-to-cloud resilience translates into improved real-world decision performance.

Conclusion

This study developed a conceptual framework for resilient edge-to-cloud AI architectures supporting distributed real-time decision making. Through synthesis of the supplied literature, the analysis established that resilience extends beyond infrastructure availability and includes data quality, uncertainty, model reliability, communication continuity, computational flexibility, privacy, and human decision support.

The proposed six-layer architecture—data acquisition, edge intelligence, adaptive orchestration, cloud intelligence, resilience management, and decision feedback—provides a structured mechanism for distributing AI workloads according to latency, confidence, resource availability, and decision requirements. Edge nodes provide rapid local processing, while cloud infrastructure supports computationally intensive analysis and centralized coordination. Adaptive orchestration connects these environments by enabling dynamic task placement and graceful degradation.

The major contribution is therefore conceptual: resilient AI should be understood as an end-to-end property of the inference-to-decision pipeline. Ground-truth uncertainty, distorted inputs, workload conditions, information protection, and diagnostic reliability must be incorporated into architectural decisions rather than addressed independently.

Future research should implement the proposed architecture in real distributed environments and evaluate it using standardized resilience metrics. Particular attention should be given to model drift, uncertainty-aware routing, heterogeneous hardware, privacy-preserving edge inference, failure recovery, and human-centered evaluation. Such research could transform edge-to-cloud AI from a computational deployment strategy into a comprehensive framework for dependable real-time decision intelligence.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 08 Pages: 122-131
OCLC – 1368736135

ISSN-2750-1396

REFERENCES

1. C. Moore and D. Toumpanakis, "Ground truth," Radiopaedia, 2021, doi: 10.53347/rID-80984.
2. E. Bowler, P. T. Fretwell, G. French, and M. Mackiewicz, "Using deep learning to count albatrosses from space: Assessing results in light of ground truth uncertainty," Remote Sens., vol. 12, no. 12, 2020, Art. no. 2026, doi: 10.3390/rs12122026.
3. Pershin, M. Kholiavchenko, B. Maksudov, T. Mustafaev, D. Ibragimova, and B. Ibragimov, "Artificial intelligence for the analysis of workload-related changes in radiologists' gaze patterns," IEEE J. Biomed. Health Inform., vol. 26, no. 9, pp. 4541–4550, Sep. 2022, doi: 10.1109/JBHI.2022.3183299.
4. L. Chang, C. Fu, Z. Wu, W. Liu, and S. Yang, "Data-driven analysis of radiologists' behavior for diagnosing thyroid nodules," IEEE J. Biomed. Health Inform., vol. 24, no. 11, pp. 3111–3123, Nov. 2020, doi: 10.1109/JBHI.2020.2969322.
5. M. Z. C. Azemin, R. Hassan, M. T. M. Izzuddin, and M. A. M. Ali, "COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: Preliminary findings," Int. J. Biomed. Imag., vol. 2020, 2020, Art. no. 8828855, doi: 10.1155/2020/8828855.
6. N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell, "A systematic review of fatigue in radiology: Is it a problem?," Amer. J. Roentgenol., vol. 210, no. 4, pp. 799–806, 2018, doi: 10.2214/AJR.17.18613.
7. S. Devi, M. N. Sahoo, K. Muhammad, W. Ding, and S. Bakshi, "Hiding medical information in brain MR images without affecting accuracy of classifying pathological brain," Future Gener. Comput. Syst., vol. 99, pp. 235–246, 2019, doi: 10.1016/j.future.2019.01.047.
8. S. Devi, S. Bakshi, and M. N. Sahoo, "Effect of situational and instrumental distortions on the classification of brain MR images," Biomed. Signal Process. Control, vol. 79, 2023, Art. no. 104177, doi: 10.1016/j.bspc.2022.104177.
9. T. K. Behera, M. A. Khan, and S. Bakshi, "Brain MR image classification using superpixel-based deep transfer learning," IEEE J. Biomed. Health Inform., early access, Oct. 21, 2022, doi: 10.1109/JBHI.2022.3216270.
10. Ramamurthy, K. (2023). AI-Driven Test Automation Frameworks for the Modern Software Quality Engineering. International Journal of Emerging Trends in Computer Science and Information Technology, 4(4), 257-269.
11. Philip, P. G. (2024). Artificial Intelligence-Driven Project Risk Prediction Models: Enhancing Decision-Making Accuracy in Large-Scale Infrastructure Projects . American Journal of Technology, 3(1), 52–69. https://doi.org/10.58425/ajt.v3i1.571
12. 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.