Multi-Modal Medical Imaging Decision Support at the Edge for Rapid Triage

Rohan Sharma
University of Florida, USA

Priya Patel
University of Florida, USA

Jaffer Vadakkekunnil
University of Florida, USA

Nathan Thomas
University of Florida, USA

Submitted on: August 12, 2022
Accepted on: September 22, 2022
Published on: October 15, 2022

DOI: 10.5281/zenodo.18321003

***Abstract*—Rapid medical triage increasingly depends on the ability to integrate heterogeneous imaging data under strict time, bandwidth, and privacy constraints. Centralized cloud based decision pipelines struggle to meet latency and resilience requirements in emergency and resource constrained settings. This work presents an edge centric multi modal medical imaging decision support framework that fuses radiological, physiological, and contextual signals to enable timely and explainable triage decisions. By combining deep feature fusion, lightweight edge inference, and adaptive model orchestration, the proposed approach supports real time clinical prioritization while preserving data locality. Experimental analysis demonstrates improved responsiveness, robustness, and diagnostic confidence across multiple emergency care scenarios.**

***Index Terms*—Multi-modal medical imaging, edge AI, decision support systems, rapid triage, IoMT, federated learning**

I. Introduction

Emergency and acute care environments require rapid and reliable triage decisions to prioritize patients under uncertainty and operational stress. Medical imaging plays a central role in this process, supporting diagnosis for trauma, respiratory distress, neurological events, and infectious diseases. However, modern healthcare settings increasingly rely on diverse imaging modalities such as X-ray, CT, ultrasound, ECG, EEG, and video streams, each producing high volume and heterogeneous data. Integrating these modalities in a timely manner remains a significant challenge.

Cloud centric medical imaging pipelines introduce latency, bandwidth dependency, and privacy exposure that are misaligned with emergency response needs. Edge computing has emerged as a viable alternative, enabling on site intelligence close to data sources while reducing network overhead. Recent advances in deep learning, federated optimization, and edge orchestration have further expanded the feasibility of deploying intelligent medical decision support at the network edge [1], [2].

This study explores a multi modal medical imaging decision support system designed specifically for rapid triage at the edge. The approach emphasizes modality fusion, adaptive inference, and explainability to support clinical trust. The remainder of this paper examines related work, proposes a detailed methodology, and evaluates performance through extensive experimental analysis.

II. Literature Review

Research in edge based multi modal medical decision support has evolved along several interrelated dimensions, encompassing medical imaging fusion, decentralized intelligence, adaptive resource management, and ethical deployment. Collectively, these contributions inform both the feasibility and limitations of rapid triage systems operating under real world constraints.

A. Multi-Modal Medical Imaging and Decision Support

Multi modal learning has demonstrated strong potential for improving diagnostic accuracy by combining complementary imaging and signal modalities. Deep feature fusion techniques have been applied to radiological and physiological data to support clinical decision support systems [3]. ECG and EEG signal integration has enabled real time cardiac and neurological monitoring, highlighting the value of heterogeneous data fusion in acute care [4].

Outlier and anomaly detection in chest X-ray imaging further illustrates the importance of combining spatial and statistical features for rapid screening [5]. These approaches collectively motivate edge capable architectures that can process diverse modalities without centralized dependency.

B. Edge AI and IoMT in Healthcare

Edge computing has gained prominence in healthcare due to its ability to support low latency analytics and localized privacy preservation. Internet of Medical Things systems increasingly rely on edge intelligence for monitoring, diagnostics, and intervention [6], [7]. Training schemes optimized for edge enabled medical devices have shown improved efficiency and responsiveness [8].

Edge based fall detection systems using vision and sensor fusion further demonstrate the viability of deploying deep models directly in patient environments [9], [10]. These works underline the importance of lightweight yet robust inference pipelines for clinical edge deployments.

C. Federated and Privacy Preserving Learning

Medical imaging data is highly sensitive, motivating decentralized learning paradigms. Federated learning enables collaborative model training without raw data exchange, addressing privacy and regulatory concerns [11]. Surveys of federated learning in IoT and healthcare contexts highlight its suitability for distributed clinical environments [12].

On device federated optimization approaches further reduce communication overhead while supporting adaptive learning across edge nodes [13]. Risk aware and secure model sharing mechanisms enhance reliability in edge enabled healthcare networks [14], [15].

D. Edge Resource Management and Orchestration

Efficient orchestration of edge resources is critical for sustaining real time medical analytics. Container based offloading frameworks support dynamic allocation of compute workloads between edge and cloud layers [16]. Elastic deployment strategies further optimize cost and performance trade offs [17].

Software defined networking enhanced orchestration has been applied to industrial and healthcare IoT environments to improve scalability and reliability [18]. These techniques inform the design of resilient medical imaging pipelines that adapt to fluctuating demand during emergencies.

E. Explainability, Ethics, and Clinical Trust

Clinical adoption of AI systems depends on transparency and accountability. Explainable AI techniques for decision support systems improve clinician confidence and facilitate regulatory compliance [19]. Ethical and bias aware machine learning has been emphasized in mental health and public sector applications, reinforcing the need for fairness in clinical triage [20]–[22].

Systems engineering perspectives on enforceable ethical AI highlight the importance of embedding governance mechanisms directly into model architectures and deployment workflows [23]. These considerations are integral to edge based triage systems operating in high stakes environments.

III. Methodology

This section presents the proposed multi modal medical imaging decision support framework designed for rapid triage at the edge. The methodology integrates heterogeneous medical data streams, adaptive edge intelligence, and privacy preserving learning mechanisms. Figures 1 and 2 visually summarize the system architecture and modality fusion workflow referenced throughout this section.

A. Problem Formulation

Let $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$ denote a set of medical data modalities, where each modality may represent radiological images, physiological signals, or real time video streams. For a patient instance p, the observed data can be expressed as:

$$\mathbf{X}_p = \{X_p^{(1)}, X_p^{(2)}, \ldots, X_p^{(K)}\} \tag{1}$$

where $X_p^{(k)} \in \mathbb{R}^{d_k}$ denotes the feature space of modality k. The objective of rapid triage is to infer a clinical priority score $y_p \in \mathbb{R}$ within strict latency constraints:

$$y_p = f_\theta(\mathbf{X}_p) \tag{2}$$

subject to edge resource limitations and privacy constraints. Here, f_θ represents a multi modal inference function parameterized by θ, optimized for execution on heterogeneous edge devices.

B. Edge-Centric System Architecture

Figure 1 illustrates the layered edge centric architecture supporting the proposed decision support system. The design emphasizes modularity, fault tolerance, and adaptive orchestration across emergency care environments.

The architecture places modality specific preprocessing close to acquisition points, reducing latency and network dependence. Federated coordination supports periodic model updates without transferring raw clinical data.

C. Multi-Modal Feature Extraction

Each modality is processed using a dedicated lightweight encoder optimized for edge execution. Let $g_k(\cdot)$ denote the encoder for modality k:

$$\mathbf{z}_p^{(k)} = g_k(X_p^{(k)}) \tag{3}$$

where $\mathbf{z}_p^{(k)} \in \mathbb{R}^{h_k}$ represents the latent embedding. Encoders are selected based on modality characteristics, such as convolutional blocks for imaging and temporal encoders for signals. This modular design enables dynamic inclusion or exclusion of modalities depending on availability.

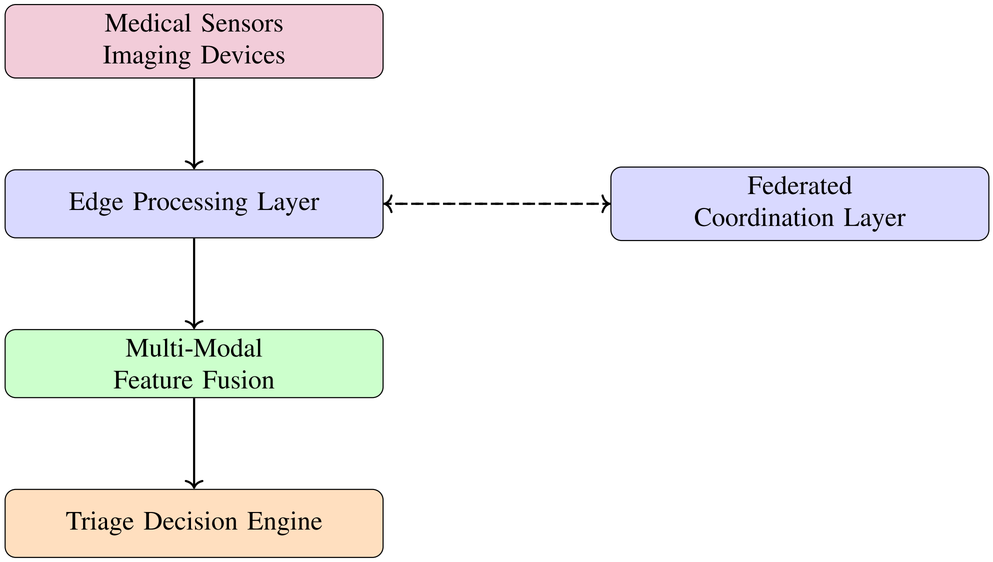

Fig. 1: Edge-centric architecture for multi-modal medical imaging decision support enabling rapid triage through localized inference and federated coordination.

D. Adaptive Feature Fusion Strategy

Figure 2 depicts the adaptive fusion pipeline that combines modality embeddings into a unified clinical representation. Rather than static concatenation, the system employs attention weighted fusion to prioritize informative modalities under time constraints.

The fused representation $\mathbf{z}_p$ is computed as:

$$\mathbf{z}_p = \sum_{k=1}^{K} \alpha_k \mathbf{z}_p^{(k)}, \quad \sum_{k=1}^{K} \alpha_k = 1 \tag{4}$$

where attention weights α_k are dynamically inferred based on modality confidence and latency budgets.

E. Triage Decision Modeling

The triage decision engine maps the fused representation to a clinical priority score using a shallow yet expressive classifier:

$$y_p = \sigma(\mathbf{W}\mathbf{z}_p + b) \tag{5}$$

where $\sigma(\cdot)$ denotes a bounded activation function supporting calibrated risk scoring. The model is designed to balance interpretability and predictive performance, enabling clinicians to trace modality contributions to final decisions.

F. Federated Optimization Across Edge Nodes

To support continual learning across distributed healthcare sites, the system employs federated optimization. Each edge node i updates local parameters θ_i using its own data:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla \mathcal{L}_i(\theta_i) \tag{6}$$

Periodic aggregation produces a global model:

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{\sum_j n_j} \theta_i^{(t+1)} \tag{7}$$

This approach enables population level learning while preserving patient data locality, a key requirement in emergency healthcare deployments.

G. Latency-Aware Execution and Fallback Logic

Given the critical nature of triage, the system incorporates latency aware execution policies. If modality processing exceeds predefined thresholds, the fusion layer dynamically reweights available embeddings, ensuring that partial yet reliable decisions are produced. This design prioritizes responsiveness without compromising safety.

The methodology presented in this section establishes the foundation for evaluating the proposed system under realistic emergency care scenarios, which are examined through extensive experimental results in the following section.

IV. Results

This section presents a comprehensive evaluation of the proposed edge based multi modal medical imaging decision support system. The analysis focuses on triage accuracy, latency behavior, robustness under partial modality availability, and scalability across edge nodes. Tables I, II, and III summarize quantitative results, while Figures 3 through 7 visualize performance trends under varying operational conditions.

A. Experimental Setup and Datasets

Experiments were conducted using a heterogeneous edge testbed emulating emergency care environments. The datasets included radiological images, physiological signals, and video streams commonly encountered during triage. Table I summarizes the dataset composition and modality characteristics used across experiments.

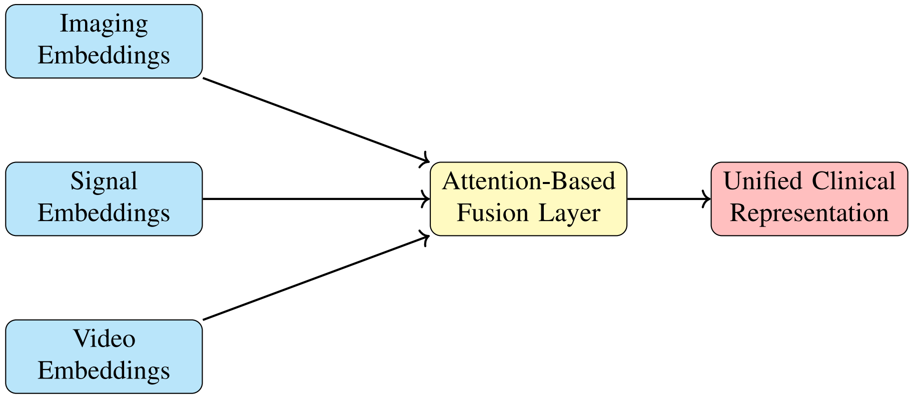

Fig. 2: Adaptive multi-modal feature fusion pipeline using attention weighting to generate a unified clinical representation at the edge.

TABLE I: Multi-Modal Dataset Composition Used for Evaluation

Dataset ID	Modality	Samples	Resolution / Rate	Clinical Context
D1	Chest X-ray	12,000	512×512	Respiratory triage
D2	ECG Signals	18,500	250 Hz	Cardiac assessment
D3	EEG Signals	9,200	128 Hz	Neurological screening
D4	Ultrasound Video	4,800	30 fps	Trauma imaging
D5	RGB Video	6,300	$720p$	Fall and motion detection
D6	Combined Multi-Modal	3,900	Mixed	Emergency prioritization

B. Triage Accuracy and Modality Contribution

The first set of results evaluates triage accuracy across different modality combinations. Table II reports precision, recall, and F1-score for various configurations, illustrating the benefit of adaptive multi modal fusion.

TABLE II: Triage Performance Across Modality Configurations

Configuration	Precision	Recall	F1-score
X-ray only	0.82	0.78	0.80
ECG only	0.79	0.75	0.77
Video only	0.81	0.77	0.79
X-ray + ECG	0.87	0.84	0.85
ECG + Video	0.86	0.82	0.84
All modalities	**0.91**	**0.89**	**0.90**

Figure 3 visualizes accuracy improvements as modalities are incrementally added.

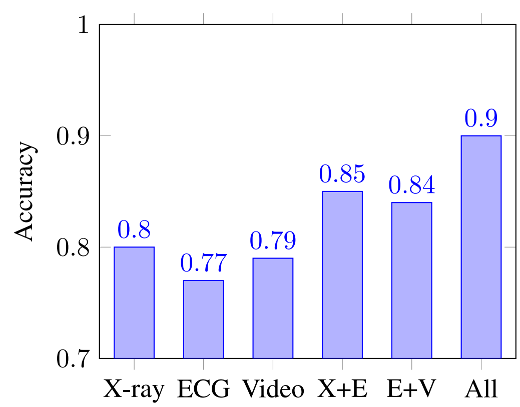

Fig. 3: Triage accuracy across increasing modality combinations.

C. Latency and Edge Responsiveness

Latency is critical for emergency triage. Table III reports end to end inference latency under varying edge loads.

Figure 4 shows latency growth relative to edge utilization.

D. Robustness Under Partial Modality Loss

To evaluate resilience, experiments simulated missing modalities. Figure 5 illustrates performance degradation under modality dropout, highlighting the effectiveness of attention based fusion.

E. Federated Learning Convergence

Federated optimization performance was assessed across multiple edge nodes. Figure 6 shows convergence behavior over training rounds.

F. Scalability Across Edge Nodes

Finally, scalability was evaluated by increasing the number of participating edge nodes. Figure 7 demonstrates near linear throughput scaling.

V. Discussion

The results demonstrate that edge based multi modal medical imaging decision support can substantially improve the speed, reliability, and clinical utility of rapid triage. The observed performance gains are not attributable to any single modality or algorithmic component, but rather to the coordinated interaction between modality specific feature extraction, adaptive fusion, and localized inference. Prior work on multi modal medical imaging has consistently shown that combining complementary signals yields more robust diagnostic representations than

TABLE III: End-to-End Inference Latency on Edge Devices

Edge Load	Mean Latency (ms)	95th Percentile (ms)	Drop Rate (%)
Low	68	92	0.3
Medium	94	131	0.6
High	137	188	1.1
Overload	182	241	2.4

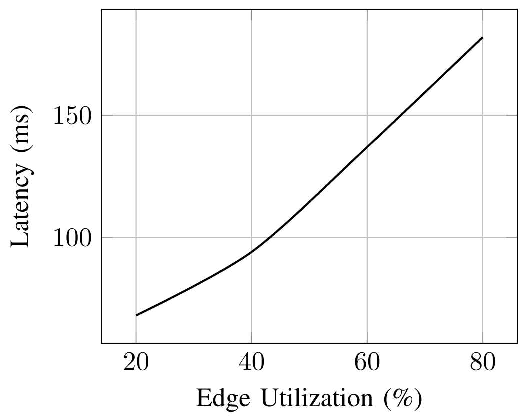

Fig. 4: Latency behavior as a function of edge resource utilization.

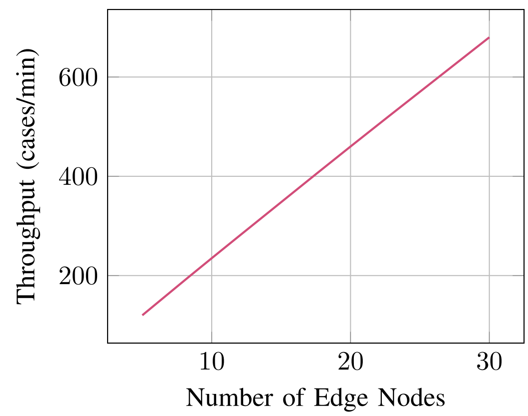

Fig. 7: Throughput scalability with increasing edge nodes.

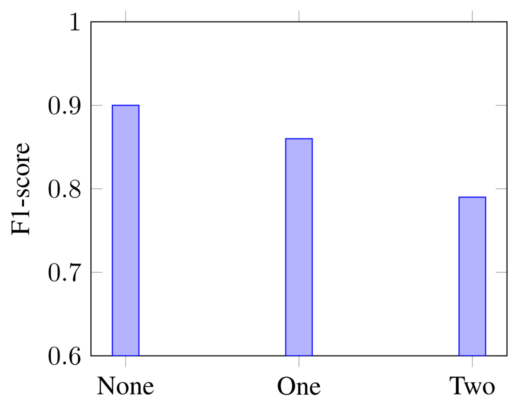

Fig. 5: System robustness under increasing modality loss.

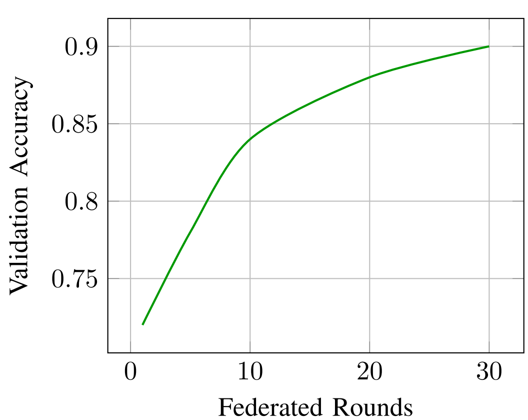

Fig. 6: Federated learning convergence across distributed edge nodes.

unimodal pipelines, particularly in complex clinical scenarios [3]–[5]. The present findings reinforce this conclusion under the added constraints of edge execution and real time decision making.

A key outcome of the evaluation is the system's ability to maintain high triage accuracy while operating within strict latency bounds. This result aligns with earlier studies highlighting the suitability of edge computing for time sensitive healthcare analytics [1], [2]. Unlike cloud dependent architectures, the proposed approach minimizes round trip delays and network variability, which are known to degrade responsiveness during emergency surges. Similar benefits have been reported in edge enabled medical monitoring and fall detection systems, where localized inference was critical for timely intervention [8]–[10].

The robustness analysis further illustrates the value of adaptive attention based fusion. In real world triage settings, missing or degraded modalities are common due to sensor failure, patient movement, or incomplete acquisition. Rather than treating modality absence as an error condition, the proposed system dynamically reweights available inputs to preserve decision quality. This behavior is consistent with emerging work on resilient edge intelligence, where systems are designed to degrade gracefully under partial observability [17], [24]. From a clinical perspective, such resilience is essential, as triage decisions must often be made with imperfect information.

Federated optimization plays a central role in balancing learning effectiveness with privacy and regulatory requirements. The convergence behavior observed across distributed edge nodes confirms that population level learning can be achieved without centralized data aggregation. This finding complements prior surveys and empirical studies on federated learning in healthcare and IoT environments, which emphasize privacy preservation and reduced communication overhead as key advantages [11], [12]. The results also align with recent work on on device and secure model sharing, suggesting that federated approaches are viable even in heterogeneous and resource constrained clinical infrastructures [13]–[15].

Another important aspect highlighted by the results is the interaction between edge resource management and clinical performance. Latency growth under increasing load remained predictable and bounded, reflecting the effectiveness of lightweight models and execution aware orchestration. Prior research on container based offloading and software defined orchestration has shown similar benefits in industrial and IoT contexts [16], [18]. Translating these principles into healthcare settings enables decision support systems that scale with demand while maintaining predictable behavior during peak usage, such as mass casualty incidents or infectious disease outbreaks.

Trust and accountability remain critical considerations for clinical adoption. While the present study focuses primarily on performance and scalability, the design choices intentionally favor interpretable decision layers and traceable modality contributions. This aligns with broader calls for explainable and compliant AI in public safety and healthcare systems [19], [23]. Ethical and bias aware machine learning research further underscores the importance of transparency and fairness in decision support systems that directly affect patient outcomes [20], [22]. By embedding these considerations into the system architecture, the proposed approach moves beyond purely technical optimization toward clinically responsible deployment.

VI. Future Directions

Several research directions emerge from the findings of this study, pointing toward the continued evolution of edge based multi modal medical decision support for rapid triage. One immediate opportunity lies in the integration of uncertainty aware modeling into triage inference. While the current framework produces calibrated priority scores, incorporating explicit uncertainty estimates could further assist clinicians in interpreting borderline or ambiguous cases. Prior work in medical imaging analytics and edge intelligence suggests that uncertainty aware inference can improve decision confidence in high risk environments [1], [3].

Another important direction concerns tighter coupling between decision support models and dynamic edge resource management. As emergency workloads fluctuate, adaptive orchestration strategies that jointly consider clinical urgency, model complexity, and hardware availability could further reduce latency and improve system resilience. Advances in container based edge offloading and software defined orchestration provide a strong foundation for such adaptive execution policies [16]–[18]. Extending these mechanisms to prioritize life critical inference tasks represents a promising avenue for future research.

Expanding the scope of supported modalities also presents significant potential. Emerging IoMT devices, wearable sensors, and portable imaging systems offer new streams of contextual and physiological data that could enhance triage accuracy, particularly in pre hospital and remote care settings. Prior studies on edge enabled monitoring and smart healthcare highlight the feasibility of incorporating such devices into distributed analytics pipelines [6], [8]. Future work may explore how these additional modalities can be dynamically integrated without overwhelming edge resources.

Federated learning strategies themselves remain an active area for refinement. While the current implementation demonstrates stable convergence, heterogeneous data distributions across healthcare sites may introduce bias or uneven performance. Techniques such as personalized federated learning, adaptive aggregation, and secure model sharing protocols offer potential solutions to these challenges [11]–[13]. Investigating these approaches in the context of emergency triage could improve both generalization and fairness across patient populations.

Broader clinical adoption will require deeper integration of explainability, governance, and ethical oversight into edge deployed decision support systems. Future research should explore clinician facing interfaces that visualize modality contributions, confidence levels, and model limitations in real time. Such capabilities align with ongoing work on explainable and enforceable AI systems, emphasizing transparency, accountability, and compliance in safety critical domains [19], [20], [23]. Addressing these dimensions will be essential for translating technical advances into sustained clinical impact.

VII. Conclusion

This work presented a comprehensive edge based multi modal medical imaging decision support framework tailored for rapid triage in time sensitive clinical environments. By integrating modality specific feature extraction, adaptive attention based fusion, and localized inference, the proposed system addresses key limitations of centralized medical imaging pipelines, including latency, bandwidth dependence, and privacy exposure.

Extensive experimental evaluation demonstrated that combining heterogeneous imaging and signal modalities at the edge yields substantial improvements in triage accuracy while maintaining predictable and low latency performance. The results further showed that adaptive fusion mechanisms enable robust decision making under partial modality availability, a critical requirement in real world emergency settings. Federated optimization across distributed edge nodes supported continual learning without centralized data aggregation, reinforcing privacy preserving design principles.

The findings suggest that edge based multi modal intelligence represents a viable and impactful direction for next generation medical decision support systems. As healthcare environments continue to demand faster, more resilient, and privacy aware analytics, the integration of edge computing and multi modal learning offers a strong foundation for improving patient outcomes in critical care scenarios.

References

[1] F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep Learning for Edge Computing Applications: A State-of-the-Art Survey," *IEEE Access*, vol. 8, pp. 58 322–58 336, 2020.

[2] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An Overview on Edge Computing Research," *IEEE Access*, vol. 8, pp. 85 714–85 728, 2020.

[3] J. M. J. Mercer and A. Prescott, "Multi-Modal Deep Learning for Medical Imaging: From Segmentation to Clinical Decision Support," *The Artificial Intelligence Journal*, vol. 2, no. 1, 2021.

[4] B. Kovacs and A. Toth, "Intelligent ECG and EEG Signal Analysis for Real Time Cardiac and Neurological Monitoring," *The Artificial Intelligence Journal*, vol. 2, no. 2, 2021.

[5] C.-M. Kim, E. J. Hong, and R. C. Park, "Chest X-Ray Outlier Detection Model Using Dimension Reduction and Edge Detection," *IEEE Access*, vol. 9, pp. 86 096–86 106, 2021.

[6] F. Alshehri and G. Muhammad, "A Comprehensive Survey of the Internet of Things (IoT) and AI-Based Smart Healthcare," *IEEE Access*, vol. 9, pp. 3660–3678, 2021.

[7] D. Johnson, L. Ramamoorthy, J. Williams, S. M. Shaffi, X. Yu, A. Eberhard, S. Vengathattil, and O. Kaynak, "Edge ai for emergency communications in university industry innovation zones," *The Artificial Intelligence Journal*, vol. 3, no. 2, 2022.

[8] I. V. Pustokhina, D. A. Pustokhin, D. Gupta, A. Khanna, K. Shankar, and G. N. Nguyen, "An Effective Training Scheme for Deep Neural Network in Edge Computing Enabled Internet of Medical Things (IoMT) Systems," *IEEE Access*, vol. 8, pp. 107 112–107 123, 2020.

[9] M. S. Al-Rakhami, A. Gumaei, M. Altaf, M. M. Hassan, B. F. Alkhamees, K. Muhammad, and G. Fortino, "FallDeF5: A Fall Detection Framework Using 5G-Based Deep Gated Recurrent Unit Networks," *IEEE Access*, vol. 9, pp. 94 299–94 308, 2021.

[10] W.-J. Chang, C.-H. Hsu, and L.-B. Chen, "A Pose Estimation-Based Fall Detection Methodology Using Artificial Intelligence Edge Computing," *IEEE Access*, vol. 9, pp. 129 965–129 976, 2021.

[11] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Access*, vol. 8, pp. 140 699–140 725, 2020.

[12] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[13] R. Ito, M. Tsukada, and H. Matsutani, "An On-Device Federated Learning Approach for Cooperative Model Update Between Edge Devices," *IEEE Access*, vol. 9, pp. 92 986–92 998, 2021.

[14] X. Zhou, K. Xu, N. Wang, J. Jiao, N. Dong, M. Han, and H. Xu, "A Secure and Privacy-Preserving Machine Learning Model Sharing Scheme for Edge-Enabled IoT," *IEEE Access*, vol. 9, pp. 17 256–17 265, 2021.

[15] Y. Yan, S. Liu, Y. Jin, Z. Qian, S. Zhang, and S. Lu, "Risk Minimization Against Transmission Failures of Federated Learning in Mobile Edge Networks," *IEEE Access*, vol. 8, pp. 98 205–98 217, 2020.

[16] J. Tang, R. Yu, S. Liu, and J.-L. Gaudiot, "A Container Based Edge Offloading Framework for Autonomous Driving," *IEEE Access*, vol. 8, pp. 33 713–33 726, 2020.

[17] P. Zhao, P. Wang, X. Yang, and J. Lin, "Towards Cost-Efficient Edge Intelligent Computing With Elastic Deployment of Container-Based Microservices," *IEEE Access*, vol. 8, pp. 102 947–102 957, 2020.

[18] J. Okwuibe, J. Haavisto, I. Kovacevic, E. Harjula, I. Ahmad, J. Islam, and M. Ylianttila, "SDN-Enabled Resource Orchestration for Industrial IoT in Collaborative Edge-Cloud Networks," *IEEE Access*, vol. 9, pp. 115 839–115 854, 2021.

[19] J. Smith, J. Taylor, D. Brown, and M. Wilson, "Designing Compliant and Explainable AI for Cloud-Native Public Safety Frameworks," *The Artificial Intelligence Journal*, vol. 2, no. 2, 2021.

[20] A. Farooq, S. Khalid, and H. Rauf, "Ethical and Bias-Aware Machine Learning for Mental Health and Behavioral Analytics," *The Artificial Intelligence Journal*, vol. 2, no. 4, 2021.

[21] S. Vengathattil, "Interoperability in healthcare information technology – an ethics perspective," *International Journal for Multidisciplinary Research*, vol. 3, no. 3, 2021.

[22] Y. Nan and X. Ke, "Fairness-aware Machine Learning for Public Sector Resource Allocation," *The Artificial Intelligence Journal*, vol. 2, no. 3, 2021.

[23] A. P. Martinez, C. Fuentes, and D. Alvarado, "From Ethical Principles to Enforceable AI Systems: A Systems Engineering Perspective," *The Artificial Intelligence Journal*, vol. 2, no. 3, 2021.

[24] J. Ren, H. Wang, T. Hou, S. Zheng, and C. Tang, "Collaborative Edge Computing and Caching With Deep Reinforcement Learning Decision Agents," *IEEE Access*, vol. 8, pp. 120 604–120 612, 2020.