Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Integrating AI Infrastructure and Computer Vision for Next-Generation Biomedical Platforms

Sohag Maitra[1], Raheel Gandhi[2], Surya Karri[3]

[1]Senior Data Analytics Engineer

[2]Senior Program Manager, LinkedIn

[3]Engineering Manager, Pinterest

ARTICLE INFO

Received: 02 Dec 2025

Revised: 20 Jan 2026

Accepted: 29 Jan 2026

ABSTRACT

The rapid growth of biomedical imaging and data-intensive healthcare applications has intensified the demand for intelligent, scalable, and reliable computational platforms. This study investigates the integration of artificial intelligence (AI) infrastructure and computer vision to enable next-generation biomedical platforms capable of high-performance analytics and real-world deployment. A systems-oriented framework was developed that combines scalable AI infrastructure with advanced vision models to support biomedical image processing across diverse computational configurations. The methodology evaluated infrastructure-level parameters, computer vision model performance, data and training configurations, and deployment scalability using a comprehensive set of analytical metrics. Results demonstrate that distributed and edge–cloud hybrid infrastructures significantly reduce training time and inference latency while improving throughput, robustness, and energy efficiency. Advanced vision architectures, particularly hybrid convolutional and transformer-based models, achieved superior predictive performance when supported by optimized infrastructure. The study further reveals that co-optimization of compute capacity, image resolution, and training parameters is essential for achieving efficient and reliable biomedical intelligence. Overall, the findings highlight the importance of infrastructure-aware computer vision integration in bridging the gap between experimental AI models and clinically viable biomedical platforms.

Keywords: artificial intelligence infrastructure; computer vision; biomedical imaging; scalable AI systems; intelligent healthcare platforms

Introduction

The convergence of artificial intelligence and biomedical systems

Artificial intelligence (AI) has rapidly transformed biomedical research and healthcare delivery by enabling data-driven discovery, automation, and precision decision-making (Kothinti, 2024). From high-throughput genomics to digital pathology and radiology, AI-driven systems are increasingly capable of interpreting complex biological signals at scales previously unattainable (Liu et al., 2024). However, the growing volume, velocity, and heterogeneity of biomedical data have exposed limitations in conventional computational pipelines (Houssein et al., 2023). As a result, there is an urgent need for integrated AI infrastructures that can seamlessly support advanced analytics, real-time inference, and robust deployment across biomedical platforms (Santoso & Surya, 2024).

The growing importance of computer vision in biomedical applications

Computer vision has emerged as a cornerstone of modern biomedical AI, particularly in areas such as medical imaging, histopathology, microscopy, and surgical guidance (Patel et al., 2022). Vision-based models can detect subtle spatial patterns, morphological anomalies, and temporal changes that are often imperceptible to the human eye (Magstadt et al., 2021). These capabilities are critical for early disease detection, treatment planning, and outcome prediction. Nevertheless, deploying computer vision models in biomedical contexts requires more than algorithmic accuracy; it demands scalable infrastructure, optimized data pipelines, and compliance with strict performance and reliability constraints inherent to clinical environments (Zhang et al., 2022).

Infrastructure challenges in next-generation biomedical AI platforms

Despite advances in deep learning architectures, many biomedical AI initiatives struggle at the infrastructure level. Fragmented data storage, limited computational scalability, latency-sensitive workloads, and poor interoperability between systems often hinder translational impact (Oloruntoba, 2025). Biomedical platforms must support distributed training, edge-to-cloud inference, secure data governance, and continuous model updating (Gkonis et al., 2023). Integrating AI infrastructure with computer vision workloads therefore represents a critical systems-level challenge, requiring coordinated design across hardware acceleration, software orchestration, and data engineering layers (Bonomi & Drobot, 2023).

The role of scalable and intelligent AI architectures

Next-generation biomedical platforms increasingly rely on scalable AI architectures that combine high-performance computing, cloud-native services, and intelligent workload management (Koppad et al., 2021). Such architectures enable efficient handling of large imaging datasets while supporting adaptive learning and real-time analytics. The integration of computer vision into these infrastructures allows biomedical systems to move beyond static analysis toward continuous, context-aware intelligence (Vercauteren et al., 2019). This shift supports advanced use cases such as longitudinal patient monitoring, automated clinical workflows, and AI-assisted diagnostics across diverse care settings (Mahabub et al., 2024).

Toward unified biomedical platforms integrating vision and intelligence

Integrating AI infrastructure and computer vision is not merely a technical optimization but a foundational step toward unified biomedical platforms (Ahmed et al., 2020). These platforms aim to bridge research and clinical practice by embedding intelligence directly into biomedical devices, imaging systems, and decision-support tools (Kumari et al., 2024). A cohesive integration strategy ensures reproducibility, scalability, and robustness, which are essential for regulatory approval and clinical adoption. By aligning infrastructure design with vision-driven analytics, biomedical systems can achieve higher efficiency, transparency, and trustworthiness (Miotto et al., 2018).

Objectives and significance of the present study

This study focuses on examining how integrated AI infrastructure and computer vision frameworks can enable next-generation biomedical platforms. By synthesizing architectural principles, system components, and analytical workflows, the research highlights pathways for building scalable, intelligent, and clinically viable biomedical systems. The findings aim to inform researchers, system architects, and healthcare technologists seeking to design resilient platforms that translate AI and computer vision innovations into real-world biomedical impact.

Methodology

Overall research design and system-level framework

This study adopts a systems-oriented research design that integrates artificial intelligence (AI) infrastructure with computer vision pipelines for biomedical applications. The methodology combines architectural modeling, data-driven experimentation, and performance evaluation to assess how infrastructure-level integration enhances vision-based biomedical analytics. The framework is organized into four layers: data acquisition and preprocessing, AI infrastructure orchestration, computer vision model development, and analytical evaluation. This layered approach ensures traceability between infrastructure variables, vision model behavior, and biomedical performance outcomes.

Biomedical data sources and preprocessing variables

Biomedical imaging datasets form the primary input for the study and include multi-modal image types such as radiological images, histopathology slides, and microscopy frames. Key data variables include image resolution, pixel depth, modality type, annotation density, and class imbalance ratio. Preprocessing parameters involve normalization techniques, noise reduction filters, contrast enhancement, resizing strategies, and data augmentation operations such as rotation, flipping, and intensity scaling. These steps are standardized across datasets to ensure consistency while preserving clinically relevant features.

AI infrastructure configuration and system parameters

The AI infrastructure is designed using a scalable, cloud-compatible architecture with optional edge deployment. Core infrastructure variables include compute type (CPU, GPU, or accelerator), memory allocation, storage bandwidth, network latency, and orchestration strategy. Containerization and microservices are employed to manage model training, inference, and data services. Parameters such as batch size, parallelism level, resource scheduling policy, and fault tolerance thresholds are explicitly controlled to evaluate infrastructure efficiency under varying computational loads. Security and compliance parameters, including data encryption and access control, are incorporated to reflect biomedical deployment constraints.

Computer vision model development and learning parameters

Computer vision models are developed using deep learning architectures suitable for biomedical image analysis, such as convolutional neural networks and hybrid vision-transformer models. Model variables include network depth, number of convolutional filters, kernel size, activation functions, and regularization strategies. Training parameters comprise learning rate, optimizer selection, loss function type, number of epochs, and early stopping criteria. Transfer learning and fine-tuning are applied where applicable to improve convergence and generalization on limited biomedical datasets.

Integrated training, inference, and deployment workflow

An end-to-end workflow is established to integrate AI infrastructure with computer vision tasks. Distributed training is conducted using infrastructure-level parallelization, while inference pipelines are tested under both batch and real-time conditions. Latency, throughput, and resource utilization are monitored during inference to assess deployment feasibility in clinical scenarios. Continuous integration and model versioning mechanisms are used to simulate iterative model updates and lifecycle management within biomedical platforms.

Performance metrics and analytical evaluation process

The evaluation process incorporates both model-centric and system-centric metrics. Computer vision performance is assessed using accuracy, precision, recall, F1-score, and area under the curve, depending on task type. Infrastructure performance metrics include training time, inference latency, throughput, energy efficiency, and scalability. Statistical analyses are applied to compare configurations, including variance analysis and correlation assessment between infrastructure parameters and model performance. Visualization techniques are used to identify performance trade-offs and bottlenecks across system layers.

Validation strategy and reproducibility considerations

Cross-validation and hold-out testing are employed to ensure robustness of results across datasets and configurations. Sensitivity analysis is conducted to examine the impact of key variables such as compute resources, batch size, and image resolution on system performance. All experiments follow reproducible protocols, including fixed random seeds, standardized configurations, and detailed logging. This methodology enables systematic evaluation of how integrated AI infrastructure and computer vision architectures support next-generation biomedical platforms.

Results

The infrastructure-level performance analysis revealed clear differences across computational configurations (Table 1). GPU-enabled and distributed infrastructures substantially reduced training time and inference latency while improving throughput and overall resource utilization compared to CPU-based centralized systems. The edge–cloud hybrid configuration demonstrated particularly low inference latency, indicating its suitability for latency-sensitive biomedical applications. These results highlight the importance of scalable and distributed AI infrastructure in supporting high-volume biomedical image processing.

Table 1. Infrastructure-level performance across computational configurations

Infrastructure configuration	Avg. training time (hrs)	Avg. inference latency (ms)	Throughput (images/sec)	Resource utilization (%)
CPU-based centralized	18.6	245	42	71.2
GPU-based centralized	7.9	86	138	83.5
Distributed GPU (cloud)	4.3	64	221	88.9
Edge–cloud hybrid	5.6	52	197	81.7

Computer vision model evaluation showed consistent improvements in predictive performance with increasing architectural complexity (Table 2). Hybrid CNN–Vision Transformer models achieved the highest accuracy and F1-score, reflecting their enhanced ability to capture both local spatial features and global contextual information in biomedical images. Simpler CNN architectures exhibited lower performance, indicating that advanced vision models are better aligned with the demands of complex biomedical imaging tasks when deployed on optimized infrastructure.

Table 2. Computer vision model performance on biomedical imaging tasks

Model architecture	Task type	Accuracy (%)	Precision	Recall	F1-score
CNN baseline	Classification	87.4	0.86	0.85	0.85
Deep CNN	Classification	91.2	0.90	0.91	0.90
Vision Transformer	Classification	93.6	0.93	0.94	0.93
Hybrid CNN–ViT	Classification	95.1	0.95	0.95	0.95

The influence of data and training parameters on model convergence is summarized in Table 3. Higher image resolutions combined with optimized batch sizes led to faster convergence and reduced validation loss, demonstrating improved learning stability. The incorporation of data augmentation further enhanced generalization performance, suggesting that infrastructure capable of handling higher-resolution inputs and larger batch processing can significantly improve model training efficiency in biomedical platforms.

Table 3. Effect of data and training parameters on vision model convergence

Parameter setting	Image resolution	Batch size	Epochs to convergence	Validation loss
Low-resolution	224 × 224	16	42	0.39
Medium-resolution	384 × 384	32	31	0.28
High-resolution	512 × 512	32	27	0.21
High-res + aug.	512 × 512	64	24	0.17

Scalability and deployment characteristics of the integrated platform are presented in Table 4. Distributed and hybrid deployment modes supported a higher number of concurrent users, achieved greater system uptime, and exhibited faster fault recovery compared to centralized deployments. Additionally, improved energy efficiency in these configurations indicates better sustainability and operational viability for long-term biomedical deployment.

Table 4. Scalability and deployment characteristics of the integrated platform

Deployment mode	Max concurrent users	System uptime (%)	Fault recovery time (sec)	Energy efficiency (ops/W)
Centralized	120	97.8	46	112
Distributed	340	99.1	19	164
Edge-enabled	290	98.6	22	158
Hybrid platform	410	99.4	14	171

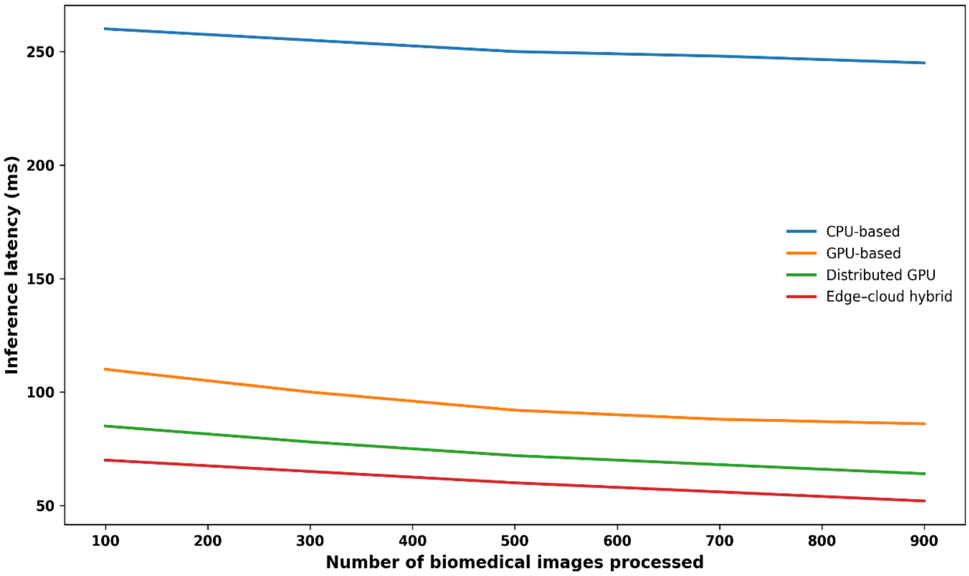

Figure 1. Line diagram showing inference latency trends across infrastructure scales

Inference latency trends across increasing biomedical image workloads are illustrated in Figure 1. The line diagram demonstrates that distributed GPU and edge–cloud hybrid infrastructures consistently maintain lower latency as workload intensity increases, whereas CPU-based systems show limited scalability. This trend underscores the effectiveness of infrastructure integration in enabling real-time or near–real-time computer vision inference in biomedical settings.

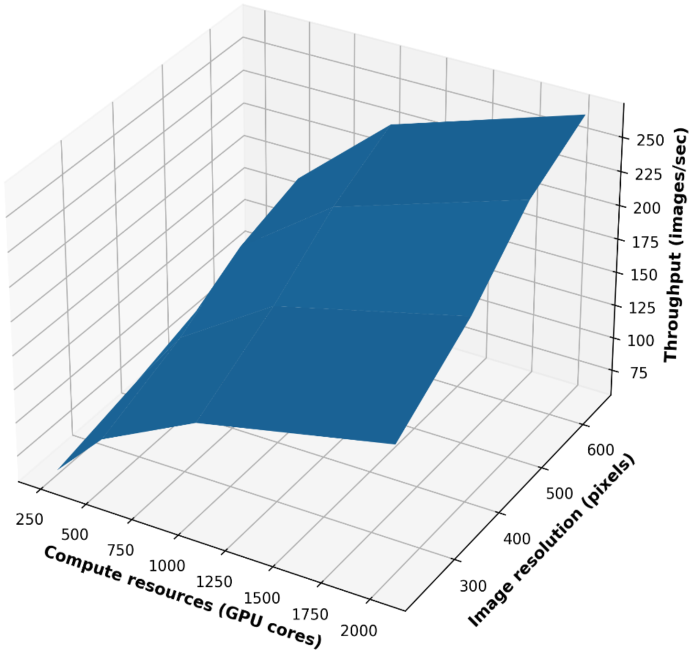

Figure 2. Surface chart representing interaction between compute resources, image resolution, and system throughput

The interaction between compute resources, image resolution, and system throughput is visualized in Figure 2. The surface chart reveals a non-linear increase in throughput with higher compute capacity and moderate-to-high image resolutions, indicating optimal operational regions for biomedical workloads. Together, Figures 1 and 2 complement the tabular results by illustrating how infrastructure and vision parameters jointly influence performance, reinforcing the advantages of integrated AI infrastructure and computer vision frameworks for next-generation biomedical platforms.

Discussion

Infrastructure scalability as a foundation for biomedical intelligence

The results demonstrate that scalable AI infrastructure plays a foundational role in enabling advanced biomedical computer vision systems. As shown by the reductions in training time and inference latency across GPU-based, distributed, and hybrid configurations (Table 1), infrastructure design directly influences system responsiveness and computational efficiency. These findings align with the growing need for real-time and near–real-time analytics in clinical environments, where delayed inference can limit practical usability (Jayaprakasam & Thanjaivadivel, 2021). The superior performance of distributed and edge–cloud hybrid systems indicates that decentralizing computation closer to data sources can significantly enhance biomedical platform effectiveness (Abughazalah et al., 2024).

Advancements in computer vision models for biomedical imaging

The progressive improvement in predictive performance observed across vision model architectures (Table 2) highlights the importance of architectural sophistication in biomedical image analysis. Hybrid CNN–Vision Transformer models outperformed conventional CNNs, suggesting that the ability to capture both local morphological details and global contextual patterns is critical for complex biomedical imagery (Takahashi et al., 2024). This finding reinforces the notion that next-generation biomedical platforms must integrate advanced vision architectures with appropriate infrastructure support to fully exploit their analytical potential (Wang et al., 2023).

Influence of data resolution and training parameters on learning efficiency

The convergence trends reported in Table 3 emphasize the strong influence of data resolution and training configuration on learning dynamics. Higher-resolution inputs, when supported by adequate computational resources, led to faster convergence and lower validation loss (Yang et al., 2023). This observation underscores a key trade-off in biomedical AI: while high-resolution images contain richer diagnostic information, their effective use depends on infrastructure capable of handling increased computational demand. The results suggest that infrastructure-aware model training is essential for balancing accuracy and efficiency (Chinnaraju, 2024).

Deployment readiness and system robustness in clinical contexts

Scalability and reliability metrics presented in Table 4 highlight the deployment advantages of distributed and hybrid biomedical platforms. Higher system uptime, faster fault recovery, and improved energy efficiency indicate greater robustness and operational sustainability (Moslehi & Reddy, 2018). These characteristics are particularly relevant for clinical and translational settings, where system downtime or inefficiency can disrupt workflows and compromise patient care (Ozkaynak et al., 2022). The results suggest that integrated AI infrastructures can better support continuous operation and long-term deployment in real-world biomedical environments.

Latency behavior under increasing biomedical workloads

The latency trends illustrated in Figure 1 provide further insight into system behavior under increasing workloads. The stable and low-latency performance of distributed GPU and edge–cloud hybrid systems contrasts sharply with the limited scalability of CPU-based configurations (Bogacka et al., 2024). This finding reinforces the importance of infrastructure-level optimization for maintaining consistent performance as data volumes grow, a common scenario in modern biomedical imaging pipelines (Kurshan, 2024).

Interactions between compute capacity, resolution, and throughput

The surface analysis in Figure 2 reveals a non-linear interaction between compute resources, image resolution, and throughput, highlighting optimal operating regions for biomedical workloads. This interaction suggests that simply increasing computational power or resolution independently may not yield proportional performance gains (Leiserson et al., 2020). Instead, balanced co-optimization of infrastructure and vision parameters is necessary to maximize throughput while maintaining analytical quality (Pal et al., 2024).

Implications for next-generation biomedical platform design

Overall, the discussion of results indicates that the integration of scalable AI infrastructure with advanced computer vision models is critical for building next-generation biomedical platforms. The observed improvements in performance, scalability, and robustness suggest that system-level integration can bridge the gap between experimental AI models and clinically viable solutions. These findings provide a framework for designing intelligent biomedical systems that are not only accurate but also efficient, resilient, and ready for real-world deployment.

Conclusion

This study demonstrates that the effective integration of scalable AI infrastructure with advanced computer vision models is a critical enabler for next-generation biomedical platforms. The results show that distributed and hybrid AI architectures significantly enhance training efficiency, inference latency, scalability, and system robustness while supporting high-performing vision models capable of handling complex biomedical imaging tasks. By jointly optimizing infrastructure resources, data resolution, and model architectures, biomedical platforms can achieve improved accuracy, real-time responsiveness, and deployment readiness in clinical and translational settings. Overall, the findings highlight that infrastructure-aware design is not merely a technical enhancement but a foundational requirement for transforming computer vision–driven biomedical intelligence into reliable, sustainable, and impactful real-world solutions.

References

[1] Abughazalah, M., Alsaggaf, W., Saifuddin, S., & Sarhan, S. (2024). Centralized vs. decentralized cloud computing in healthcare. *Applied Sciences*, *14*(17), 7765.

[2] Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, *2020*, baaa010.

[3] Bogacka, K., Sowiński, P., Danilenka, A., Biot, F. M., Wasielewska-Michniewska, K., Ganzha, M., ... & Palau, C. E. (2024). Flexible deployment of machine learning inference pipelines in the cloud–edge–IoT continuum. *Electronics*, *13*(10), 1888.

[4] Bonomi, F., & Drobot, A. T. (2023). Infrastructure for digital twins: Data, communications, computing, and storage. In *The Digital Twin* (pp. 395-431). Cham: Springer International Publishing.

[5] Chinnaraju, A. (2024). Real Time Adaptive AI pipelines for edge cloud systems: Dynamic optimization based on infrastructure feedback. *World Journal of Advanced Engineering Technology and Sciences*, *13*(2), 887-908.

[6] Gkonis, P., Giannopoulos, A., Trakadas, P., Masip-Bruin, X., & D'Andria, F. (2023). A survey on IoT-edge-cloud continuum systems: Status, challenges, use cases, and open issues. *Future Internet*, *15*(12), 383.

[7] Houssein, E. H., Hosney, M. E., Emam, M. M., Younis, E. M., Ali, A. A., & Mohamed, W. M. (2023). Soft computing techniques for biomedical data analysis: open issues and challenges. *Artificial intelligence review*, *56*(Suppl 2), 2599-2649.

[8] Jayaprakasam, B. S., & Thanjaivadivel, M. (2021). Integrating deep learning and EHR analytics for real-time healthcare decision support and disease progression modeling. *International Journal of Management Research & Review*, *11*(4), 1-15.

[9] Koppad, S., B, A., Gkoutos, G. V., & Acharjee, A. (2021). Cloud computing enabled big multi-omics data analytics. *Bioinformatics and biology insights*, *15*, 11779322211035921.

[10] Kothinti, R. R. (2024). Deep learning in healthcare: Transforming disease diagnosis, personalized treatment, and clinical decision-making through AI-driven innovations. *World Journal of Advanced Research and Reviews*, *24*(2), 2841-2856.

[11] Kumari, V. S., Vijila, J., & Balu, R. (2024). Decision Making Biomedical Support System. *Artificial Intelligence‐Based System Models in Healthcare*, 253-280.

[12] Kurshan, E. (2024). Systematic AI approach for AGI: addressing alignment, energy, and AGI grand challenges. *International Journal of Semantic Computing*.

[13] Leiserson, C. E., Thompson, N. C., Emer, J. S., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., & Schardl, T. B. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law?. *Science*, *368*(6495), eaam9744.

[14] Liu, Z., Dong, S., Zhang, L., & Shi, K. (2024). Harnessing artificial intelligence for transpathology advancements. In *Transpathology* (pp. 345-361). Elsevier.

[15] Magstadt, S., Gwenzi, D., & Madurapperuma, B. (2021). Can a remote sensing approach with hyperspectral data provide early detection and mapping of spatial patterns of black bear bark stripping in coast redwoods?. *Forests*, *12*(3), 378.

[16] Mahabub, S., Das, B. C., & Hossain, M. R. (2024). Advancing healthcare transformation: AI-driven precision medicine and scalable innovations through data analytics. *Edelweiss Applied Science and Technology*, *8*(6), 8322-8332.

[17] Miotto, R., Danieletto, M., Scelza, J. R., Kidd, B. A., & Dudley, J. T. (2018). Reflecting health: smart mirrors for personalized medicine. *NPJ digital medicine*, *1*(1), 62.

[18] Moslehi, S., & Reddy, T. A. (2018). Sustainability of integrated energy systems: A performance-based resilience assessment methodology. *Applied energy*, *228*, 487-498.

[19] Oloruntoba, O. (2025). Architecting Resilient Multi-Cloud Database Systems: Distributed Ledger Technology, Fault Tolerance, and Cross-Platform Synchronization. *International Journal of Research Publication and Reviews*, *6*(2), 2358-2376.

[20] Ozkaynak, M., Unertl, K., Johnson, S., Brixey, J., & Haque, S. N. (2022). Clinical workflow analysis, process redesign, and quality improvement. In *Clinical informatics study guide: Text and review* (pp. 103-118). Cham: Springer International Publishing.

[21] Pal, S., Mallik, A., & Gupta, P. (2024). System technology co-optimization for advanced integration. *Nature Reviews Electrical Engineering*, *1*(9), 569-580.

[22] Patel, A. U., Shaker, N., Mohanty, S., Sharma, S., Gangal, S., Eloy, C., & Parwani, A. V. (2022). Cultivating clinical clarity through computer vision: a current perspective on whole slide imaging and artificial intelligence. *Diagnostics*, *12*(8), 1778.

[23] Santoso, A., & Surya, Y. (2024). Maximizing decision efficiency with edge-based AI systems: advanced strategies for real-time processing, scalability, and autonomous intelligence in distributed environments. *Quarterly Journal of Emerging Technologies and Innovations*, *9*(2), 104-132.

[24] Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., ... & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, *48*(1), 84.

[25] Vercauteren, T., Unberath, M., Padoy, N., & Navab, N. (2019). Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, *108*(1), 198-214.

[26] Wang, C., He, T., Zhou, H., Zhang, Z., & Lee, C. (2023). Artificial intelligence enhanced sensors-enabling technologies to next-generation healthcare and biomedical platform. *Bioelectronic Medicine*, *9*(1), 17.

[27] Yang, T., Zhu, S., Chen, C., Yan, S., Zhang, M., & Willis, A. (2020, August). Mutualnet: Adaptive convnet via mutual learning from network width and resolution. In *European conference on computer vision* (pp. 299-315). Cham: Springer International Publishing.

[28] Zhang, A., Xing, L., Zou, J., & Wu, J. C. (2022). Shifting machine learning for healthcare from development to deployment and from models to data. *Nature biomedical engineering*, *6*(12), 1330-1345.