International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.1 (2026) pp. 141-148

http://www.ijcesen.com

ISSN: 2149-9144

Research Article

DevOps for Diagnostics: Automated Continuous Integration/Continuous Deployment (CI/CD) of Multi-Modal AI Models into Hospital PACS and LIS

Ashwini Pankaj Mahajan*

Independent Researcher, USA

* **Corresponding Author Email:** ashwinipankajmahajan@gmail.com - **ORCID:** 0000-0002-3337-7850

Article Info:

DOI: 10.22399/ijcesen.4713
Received : 08 November 2025
Revised : 29 December 2025
Accepted : 03 January 2026

Keywords

Continuous Integration/Continuous Deployment,
Healthcare AI Operations,
Multi-Modal Diagnostics,
Cloud-Native Architecture,
Clinical Workflow Integration

Abstract:

Diagnostic healthcare has been confronted with a consistent problem of deploying multi-modal models of artificial intelligence into the current systems of Picture Archiving and Communication Systems (PACS) and Laboratory Information Systems (LIS) because of the long-term implementation cycle, version management difficulties, and workflow interruptions. Continuous Integration/Continuous Deployment pipelines based on DevOps are radically innovative in that they automate the entire lifecycle of diagnostic AI models, starting with their creation up to their on-production maintenance. Kubernetes-based cloud-native architectures have the scalability, resilience, and compute efficiency required to handle computationally intensive diagnostic loads, as well as to regulate and govern data. The proposed framework will enforce automatic model versioning, drift detection policies, multi-phase validation policies, and a smooth integration with hospital enterprise systems via DICOM and HL7 interfaces. Empirical assessment in various healthcare facilities proves significant success rates in deployments, system stability, and system processing time without affecting clinical operations and diagnostic results during automated retraining loops. Incremental deployment models, such as canary releases and full monitoring infrastructure, allow safe updating of models without interfering with running clinical processes. Federated learning has the capability of integrating to enable the multi-institutional models to become improved, coupled with privacy preservation and regulation limitations on data sharing. Clinical acceptance evaluations indicate that there is little workflow interference and a gradual increase in user confidence due to the coherent model activity and dependability. The framework is effective in considering the main issue of long-term effectiveness because over time, the population of patients, the prevalence of the disease, and clinical practices may change. Computerized governance protocols and model registry centralized offer institutional supervision and decrease the IT support load, and facilitate the scalable implementation of AI-powered diagnostics in heterogeneous healthcare settings.

1. Introduction

Multi-modal artificial intelligence is expected to gain importance in the field of healthcare diagnostics to support clinical decision-making in imaging, pathology, and genomic analysis. The combination of DevOps principles and predictive maintenance functions became the new way to transform the healthcare IT infrastructure, where the constant monitoring and automated deployment plans are used to provide the reliability and maximum performance of such a system [1]. Conventional deployment methods have long had the effect of leading to the establishment of very long implementation cycles, as healthcare organizations have had problems balancing between innovation and business continuity. The implementation of DevOps practices specifically designed to meet the needs of healthcare facilities can help resolve such issues through introducing automated workflows covering the whole cycle of AI model deployment, including the initial creation, as well as production maintenance and eventual decommissioning. The convergence of cloud-native architectures with healthcare data analytics has fundamentally altered the landscape of diagnostic AI deployment. Scalable and secure healthcare data analytics solutions based on cloud-native underpinnings offer the infrastructure required to run computationally intensive diagnostic models and, at the same time, meet the rigid regulatory demands [2]. These architectures allow healthcare facilities to take advantage of distributed computing facilities, employing powerful protection protocols, and guarantee the control of data in multi-modal diagnostic structures. A move towards cloud-native deployment strategies is indicative of a larger realization that healthcare IT has of the fact that traditional monolithic deployment methods are insufficient to meet the dynamic needs of the modern use of AI-driven diagnostics, especially when one considers the necessity of real-time model updates, performance monitoring, and dynamically scaling deployment to clinical demand patterns.

The implementation of diagnostic AI has transformed the state of affairs radically due to the integration of cloud-native solutions with healthcare data analytics. The infrastructure needed to operate computationally-intensive diagnostic models with compliance to the requirements established by strict regulatory standards is offered in scalable and secure healthcare data analytics platforms that are based on cloud-native foundations [2]. Such architectures allow health facilities to take advantage of distributed computing facilities, put up effective security mechanisms, and guarantee data integrity across multi-modal therapeutic systems. This migration to cloud-native deployment models represents a more general acknowledgment in the sphere of healthcare IT of the inability of traditional monolithic architectures to effectively sustain the dynamic needs of the contemporary AI-driven diagnostics, especially in the context of the necessity of real-time updates to the model, live performance metrics, and scalability based on clinical trends.

2. Architectural Framework and Infrastructure Design

The proposed CI/CD framework adopts a cloud-native architecture built upon containerized microservices and orchestrated through Kubernetes clusters, leveraging the comparative advantages of container orchestration platforms for healthcare AI deployment. Cloud-native architectures employing Kubernetes demonstrate superior capabilities for managing healthcare AI workloads compared to serverless computing alternatives, particularly when handling stateful applications that require persistent connections to hospital information systems and maintaining consistent performance under variable clinical loads [3]. The architectural solution has several layers of abstraction between model execution environments and infrastructure that can support easy migration between cloud providers and between hybrid clouds, and still assist in meeting healthcare data residency requirements. Given the volumetric clinical demand patterns across the day-to-day working cycles, the Kubernetes orchestration layer offers automatic scaling features that automatically vary the amount of computational resources in relation to the incoming diagnostic study volumes to make sure that the processing capacity matches the patterns of clinical demand.

The infrastructure has four main layers that collaborate to address the entire model lifecycle. The development environment creates standardized model catalogs that have version control systems that not only monitor model artifacts but also training data, hyperparameter setups, and performance measures using a detailed metadata management framework. Container images capture the full model dependencies and would be reproducible across all development, testing, and production environments, and would remove the environment drift complications that are characteristic of healthcare AI implementations. The integration pipeline uses automated build triggers (code commits or periodic) to run a sequence of operations such as dependency resolution, model compilation, unit testing, and container image construction. These processes are monitored by Continuous Integration servers, which use real-time telemetry to generate detailed logs and notify development teams of failures at any point in the process without losing full audit trails necessary to meet the regulatory compliance documentation requirements.

Validation protocols represent an essential element that renders the implementation of AI in healthcare distinct in comparison with traditional software delivery, which involves a multi-phase evaluation that includes both technical performance indicators and clinical accuracy measures and regulatory compliance checks. The frameworks carry out automated validation processes which compare model performance to a set of held-out test data and use standard measures of diagnostic accuracy, as well as model behavior under edge cases and adversarial inputs that could happen in clinical practice. Clinical validation is a synthetic clinical scenario testing whereby the models are exposed to retrospective cases with known diagnoses and simulate real-world diagnostic workflow, and present potential failure modes before production deployment. The validation infrastructure has specially reserved testing environments that simulate production PACS and LIS configurations, allowing full integration testing without risking disrupting live clinical processes. Healthcare data handling standards, medical device software standards, and interoperability standards are enforced through regulatory compliance checks, which are automated to generate validation documentation needed in regulatory submissions and institutional review procedures.

The production deployment interface manages the critical transition from validated models to active clinical systems through sophisticated orchestration mechanisms that minimize deployment risk. Rather than implementing direct replacement of existing models, the framework employs progressive delivery strategies, including canary deployments, where new model versions initially process a carefully controlled subset of incoming diagnostic studies. Monitoring at Canary phases takes action on several metrics, such as inference latency, distributions of prediction confidence, error rates, and resource usage by the system, to identify early indicators of unexpected behaviors before full deployment. The interface has a two-way communication interface with both PACS and LIS by use of standard protocols like DICOM (medical imaging) and HL7 (laboratory data) to have message queuing and re-try logic in order to deliver messages reliably even when the network is unavailable, or the system is undergoing maintenance. The abstraction layers of API gateways allow model endpoints to be independent of hospital information systems, allowing model updates where clinical applications do not need to change, as well as A/B testing and staged rollout approaches that mitigate the risk of deployment in a heterogeneous clinical environment.

3. Model Lifecycle Management and Automated Retraining

Lifecycle management is not limited to initial deployment, since it includes continuous monitoring, performance analysis, and updating the model in accordance with the changing clinical data trends. The framework applies extensive data drift detection systems that constantly demonstrate the evaluation of incoming diagnostic data and model outputs in order to establish the presence of distributional changes that can undermine the model accuracy. Data drift detection and mitigation strategies employ statistical process control methods combined with machine learning-based anomaly detection to identify subtle changes in data characteristics that precede observable performance degradation [5]. The monitoring infrastructure tracks multiple drift indicators, including feature distribution shifts, prediction confidence trends, and outcome correlations, aggregating these signals into composite health metrics that trigger automated responses when crossing predefined thresholds. This proactive approach to drift detection enables healthcare organizations to maintain model performance through timely retraining interventions rather than reactive corrections after accuracy degradation becomes clinically apparent.

Automated retraining protocols activate when monitoring systems detect performance metrics falling below acceptable thresholds or when drift detection algorithms identify significant distributional changes in clinical data streams. These protocols retrieve updated training datasets from hospital data warehouses, implementing sophisticated data curation strategies that maintain appropriate class balance, representation of rare conditions, and patient privacy protections throughout the retraining process. The retraining pipeline incorporates federated learning capabilities that enable multi-institutional model improvement without centralizing sensitive patient data, addressing both performance enhancement objectives and regulatory constraints on data sharing. Transfer learning algorithms initialize new model versions with the weights of a previous iteration, cutting the computational needs and training time of model training significantly, and retaining learned representations that are also applicable in later generations of models. To support regulatory auditing and institutional governance, the framework ensures there is detailed provenance tracking during retraining cycles, the source of data, training hyperparameters, training validation results, and performance benchmarks.

Version management systems maintain full model lineages, recording the evolutionary history of a model since the first deployment, in a rich metadata format that records model features, model performance measures, and deployment history. The model versions are assigned an immutable identifier that is created by cryptographic hashing of model artifacts, training configurations, and validation data, and the authenticity of a model version can be verified, and unauthorized modifications prevented. The framework adopts advanced version control procedures that, in turn, offer branching strategies on experimental model variants, facilitating parallel development of specialized models to particular clinical situations whilst keeping production deployments stable. Automated governance policies implement institutional conditions of minimum validation sample size, performance levels, and regulatory approvals for production deployment of models. Such policies can be configured using declarative specifications that represent institutional standards, and then allow healthcare organizations to tailor approval workflows to their risk tolerance and regulatory requirements.

Model registry services store centralized lists of known model versions with related metadata, performance characteristics, deployment status, and lineage information that are reached via intuitive interfaces that do not need technical knowledge on the part of clinical administrators. The registry provides advanced search and filtering features, allowing the selection of a model depending on diagnostic modality, anatomical region, performance measures, regulatory approval, or deployment history. Interaction with the overall MLOps system ensures that registry data is kept up-to-date with deployed model versions, avoiding differences between the reported capabilities and reality.

4. Integration with Hospital Enterprise Systems

Effective implementation involves smooth interoperability with the existing enterprise systems of the hospital, coupled with minimal impacts on the existing clinical operations, which is complicated by the heterogeneity of the health care IT infrastructure. Implementation of AI in the PACS settings requires a close evaluation of the DICOM-based infrastructure needs, such as adherence to the imaging communication standards, integration with radiologist workstations, and maintenance of the current image routing practices that have been refined during years of institutional utilization [7]. The framework uses DICOM-compliant middleware, which interposes imaging studies at various stages of the PACS workflow, such as modality acquisition, radiologist worklist presentation, and archive storage, which is non-blocking asynchronous, with AI model inference running parallel to normal PACS workflows. This architectural design guarantees that the processing latency of the models does not introduce clinical access latency to the images, which is very important in ensuring that radiologist productivity is not hampered and the workflow is not subjected to a bottleneck situation during high operational times.

The middleware then executes advanced message routing functionality that sends studies to the right AI models, depending on the type of examination, anatomical region, clinical indication, and institutional practices, without breaking compatibility with a variety of PACS vendor implementations.

Studies are transmitted simultaneously to conventional storage destinations and model inference endpoints through message duplication mechanisms that operate transparently within the existing DICOM network topology. Inference results return to PACS as structured reports conforming to DICOM Structured Reporting templates, enabling standardized representation of AI findings that integrate seamlessly with existing radiology information systems and clinical documentation workflows. The framework supports multiple result delivery modalities, including embedded image annotations, separate structured report objects, and integration with third-party visualization tools, providing flexibility to accommodate diverse institutional preferences and regulatory requirements regarding AI result presentation.

LIS integration follows analogous principles adapted to the distinct characteristics of laboratory data workflows, implementing HL7 messaging interfaces that subscribe to relevant message types, including laboratory orders, specimen tracking updates, and preliminary results. The framework processes laboratory data through specialized AI models designed for interpreting complex laboratory panels, identifying abnormal patterns, and flagging results requiring urgent clinical attention. Advanced deep learning approaches for clinical laboratory test interpretation have demonstrated substantial improvements in diagnostic accuracy and turnaround time, particularly for complex test panels requiring correlation across multiple analytes [8]. The integration architecture implements real-time processing pipelines that analyze laboratory results as they become available, generating interpretive reports and clinical decision support recommendations that seamlessly integrate with existing LIS workflows and physician order entry systems.

Low-latency requirements in clinical settings necessitate careful optimization of inference execution across both imaging and laboratory diagnostic domains. The framework employs multiple optimization strategies, including model quantization that reduces numerical precision while maintaining diagnostic accuracy, pruning techniques that remove redundant model parameters, and knowledge distillation that transfers capabilities from large, complex models to smaller, efficient variants suitable for real-time deployment. GPU acceleration supports parallel processing of multiple studies, with the framework implementing intelligent batch scheduling that aggregates similar studies to maximize GPU utilization while maintaining acceptable latency for individual results. Caching plans archive the results of intermediate computations on common regions of the anatomy or common diagnostic appearances, and save significant amounts of processing time on routine cases by retrieving a previously computed representation of features. The optimization infrastructure will constantly observe measurements of inference performance, and will automatically scale the batch sizes, caching parameters, and resource usage to sustain optimum throughput and latency properties as clinical load patterns change.

5. Performance Evaluation and Clinical Impact

The framework was empirically assessed by being implemented in a variety of healthcare facilities with various infrastructure layouts, clinical specialties, and patient demographics, giving an all-encompassing analysis of technical capacity and clinical workflow integration. The evaluation methodology incorporated quantitative performance metrics, including deployment success rates, model update frequencies, system reliability measurements, and processing latency distributions across different diagnostic modalities and institutional contexts. Integration of AI systems into clinical workflows presents multifaceted challenges encompassing technical infrastructure requirements, organizational change management, clinician acceptance, and alignment with existing care delivery processes [10]. The framework evaluation specifically examined these integration dimensions through structured clinician interviews, workflow observation studies, and longitudinal tracking of adoption patterns across radiology, pathology, and laboratory medicine departments.

Automation metrics of deployments showed significant progressions between manual deployment processes, and the automated CI/CD pipeline had large success rates regarding both the first deployment of the model and its future upgrades in a variety of institutional settings. The time between the releases of the model version was reduced significantly from the days when the process was done manually, and the process became more responsive to emergent diagnostic demands, performance degradation indications, and access to better model structures. The reliability metrics of the system showed that there was an outstanding uptime in the production deployments, and the extensive monitoring infrastructure facilitated quick identification and resolution of the rare malfunctions that arose. System disruption root cause analysis showed that most of the faults in the system were due to planned system downtimes and infrastructure hangovers as opposed to failure of the framework, which should have been very strong based on the quality of the architectural design and implementation, and was confirmed.

Processing latency measurements across different diagnostic modalities confirmed that inference times remained well below clinically significant thresholds throughout the evaluation period. Imaging models processing various examination types, including radiographs, CT scans, MRI studies, and ultrasound examinations, consistently delivered results within timeframes that integrated seamlessly into radiologist reading workflows. Pathology whole-slide image analysis maintained processing speeds that enabled real-time review during tumor board conferences and multidisciplinary care planning sessions. Laboratory test interpretation models provided results within windows that supported incorporation into physician rounding workflows and clinical decision-making processes. Clinical feedback consistently indicated these latencies were acceptable, with surveyed clinicians rating processing speed favorably and reporting minimal disruption to their established diagnostic practices.

Diagnostic accuracy metrics tracked across successive model versions confirmed the framework's ability to maintain stable performance through automated retraining cycles while also demonstrating gradual accuracy improvements as models incorporated increasingly diverse training data. Statistical analysis revealed that automated retraining successfully addressed detected performance drift, restoring model accuracy when monitoring systems identified degradation trends. Laboratory data interpretation models showed progressive accuracy enhancements as federated learning protocols enabled the incorporation of multi-institutional case examples, expanding model exposure to rare conditions and unusual presentations that individual institutions encounter infrequently. The framework's ability to maintain accuracy while continuously evolving models addresses a fundamental challenge in healthcare AI deployment: ensuring long-term effectiveness as patient populations, disease prevalence, equipment characteristics, and clinical practices change over time.

Clinical workflow integration assessments employed multiple methodologies, including structured interviews with radiologists, pathologists, and laboratory medical directors, direct observation of clinical workflows incorporating AI systems, and longitudinal surveys tracking clinician perceptions and acceptance over extended deployment periods. Participants consistently reported minimal disruption to established practices, attributing seamless integration to the framework's non-blocking asynchronous processing design that allows AI analysis to occur in parallel with standard clinical activities. Acceptance ratings demonstrated progressive improvement over time as clinical users developed trust in model consistency, reliability, and clinical utility through repeated positive experiences. Several institutions reported that automated model management reduced IT support burden by eliminating manual deployment coordination, troubleshooting activities, and version management overhead previously required for diagnostic AI systems, translating automated infrastructure benefits into measurable operational efficiencies and cost reductions.

Table 1: *DevOps Integration Components for Healthcare AI Infrastructure [1,2]*

Component Category	Implementation Strategy	Healthcare-Specific Consideration
Predictive Maintenance	Continuous monitoring with automated alerting	System reliability for clinical operations
Cloud-Native Analytics	Distributed computing with data governance	Regulatory compliance and scalability
Deployment Automation	Version-controlled model artifacts	Multi-modal diagnostic compatibility
Security Protocols	Encryption and access control	HIPAA compliance requirements

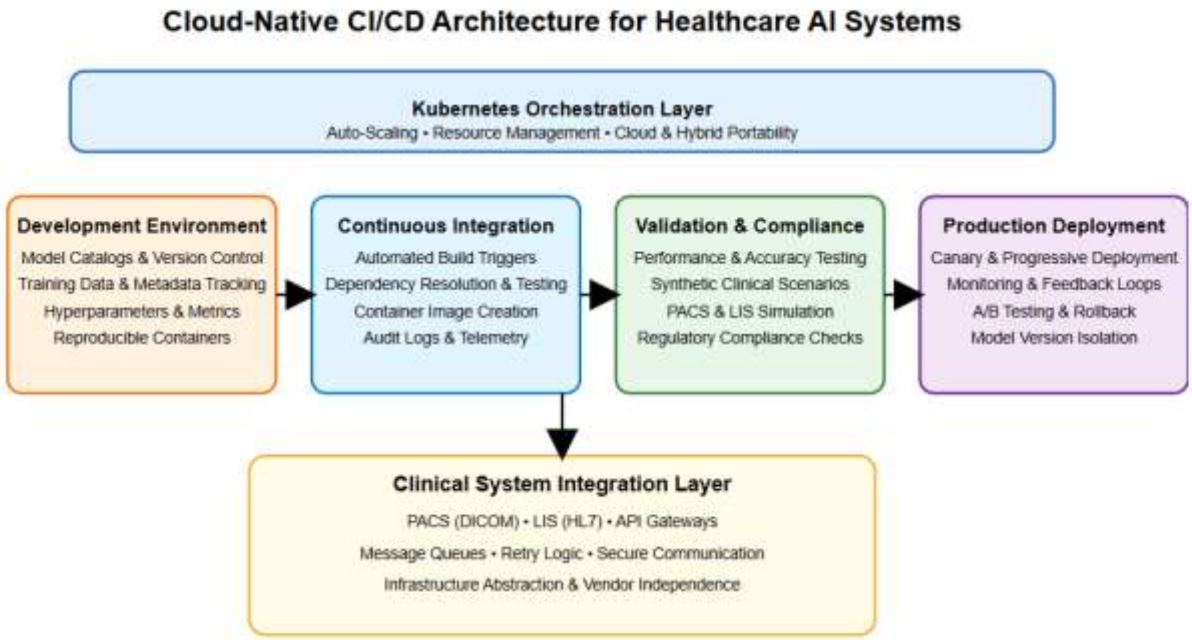

Figure 1: *Architectural Framework and Infrastructure Design*

Table 2: *Cloud Infrastructure Comparison for Healthcare AI Deployment [3,4]*

Infrastructure Aspect	Kubernetes Orchestration	Serverless Computing	Healthcare Optimization
Stateful Application Support	Persistent connections maintained	Limited state management	Critical for PACS/LIS integration
Performance Consistency	Predictable under variable loads	Variable cold-start latency	Essential for clinical workflows
Resource Scaling	Dynamic horizontal scaling	Event-driven auto-scaling	Adapts to diagnostic study volumes
Clinical Integration	Bidirectional communication support	Asynchronous processing	Maintains workflow continuity

Table 3: *Data Drift Detection and Model Lifecycle Management [5,6]*

Lifecycle Stage	Monitoring Mechanism	Automated Response	Governance Control
Performance Tracking	Statistical process control	Retraining protocol activation	Threshold-based triggers

Drift Detection	Feature distribution analysis	Dataset curation and update	Institutional validation requirements
Version Management	Cryptographic identifier generation	Immutable artifact tracking	Approval workflow enforcement
Model Registry	Metadata cataloging	Synchronization with deployments	Multi-stakeholder collaboration

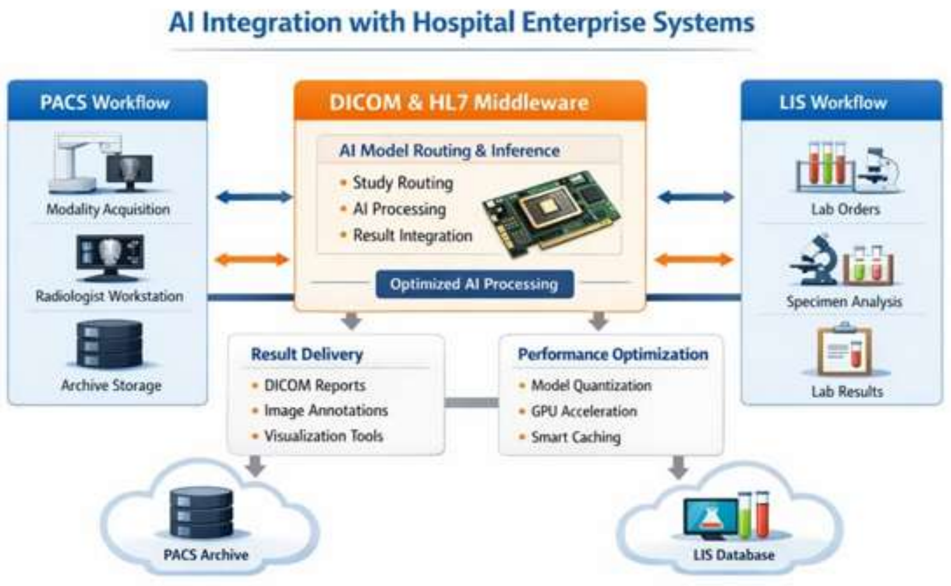

Figure 2. *Architectural Framework and Infrastructure Design for Cloud-Native Healthcare AI CI/CD Pipelines*

Table 4: *Hospital Enterprise System Integration Architecture [7,8]*

Integration Layer	PACS Implementation	LIS Implementation	Optimization Strategy
Protocol Interface	DICOM-compliant middleware	HL7 messaging subscription	Asynchronous processing
Data Routing	Multi-point interception	Real-time pipeline processing	Message duplication mechanisms
Result Delivery	Structured reporting templates	Interpretive report generation	Standardized format compliance
Performance Optimization	GPU-accelerated inference	Batch scheduling intelligence	Caching and quantization

6. Conclusions

The implementation of DevOps-driven CI/CD pipelines establishes effective mechanisms for managing the complete lifecycle of multi-modal diagnostic AI models within hospital enterprise systems. Automated operations that include version control, validation, deployment, and retraining ensure smooth integration with the current PACS and LIS infrastructure and significantly decrease deployment cycles and enhance the reliability of the system as compared to manual processes. The principles of cloud-native architecture through containerization and API-based integration are particularly highly adaptable to the healthcare AI implementation needs due to the ability to scale to the needs of the institution (including its growth), maintain isolation (including regulatory compliance), and flexibility (including the need to accommodate heterogeneous hospital IT environments). The expressed ability to preserve diagnostic accuracy by retraining models using automated mechanisms deals with the vital issue of preserving the efficiency of the models, as the population of patients, the prevalence of the disease, and practice change over time. Further directions would be towards mechanisms of model switching that do not result in any downtime to allow a full switch between model versions, and constant transitions between model versions, and provide different patients with different model versions.

Automated real-time drift detection systems employing statistical process control and change-point detection algorithms could substantially reduce the time between drift occurrence and corrective action, providing dynamic safeguards against unexpected model behaviors without human intervention. Integration of explainability mechanisms into automated deployment pipelines presents opportunities to address emerging regulatory frameworks emphasizing AI transparency through automated generation and validation of model explanations alongside predictions. Multi-institutional deployment coordination through federated learning architectures and distributed model registries could accelerate the diffusion of diagnostic innovations while maintaining local control over deployment decisions and data governance. The convergence of automated deployment infrastructure with emerging diagnostic modalities, including multi-omics integration, digital pathology, and point-of-care diagnostics, necessitates the continued evolution of CI/CD frameworks as diagnostic complexity increases and model architectures incorporate multiple data modalities simultaneously. The architectural patterns and implementation principles demonstrated position healthcare institutions to harness the full potential of AI-driven diagnostics while maintaining the reliability, safety, and workflow integration essential to clinical practice.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.

- **Author contributions:** The authors declare that they have equal right on this paper.

- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Sidhartha Velishala, "DevOps and AI for predictive maintenance in healthcare IT infrastructure," ResearchGate, 2025. [Online]. Available: https://www.researchgate.net/publication/388426240_DEVOPS_AND_AI_FOR_PREDICTIVE_MAINTENANCE_IN_HEALTHCARE_IT_INFRASTRUCTURE

[2] Rishi Kumar Sharma, "Enabling scalable and secure healthcare data analytics with cloud-native AI architectures," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/387862096_Enabling_Scalable_and_Secure_Healthcare_Data_Analytics_with_Cloud-Native_AI_Architectures

[3] Gireesh Kambala, "Cloud-Native Architectures: A Comparative Analysis of Kubernetes and Serverless Computing," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/388717188_Cloud-Native_Architectures_A_Comparative_Analysis_of_Kubernetes_and_Serverless_Computing

[4] David B. Olawade et al., "Artificial intelligence in clinical trials: A comprehensive review of opportunities, challenges, and future directions," ScienceDirect, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505625003582

[5] Naveen Kodakandla, "Data drift detection and mitigation: A comprehensive MLOps approach for real-time systems," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/388187259_Data_drift_detection_and_mitigation_A_comprehensive_MLOps_approach_for_real-time_systems

[6] Masooma Hassan, et al., "Artificial intelligence governance framework for healthcare," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/385350124_Artificial_intelligence_governance_framework_for_healthcare

[7] Dan Victor, "Deployment of AI in PACS / DICOM-based Infrastructure," ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/397614558_Deployment_of_AI_in_PACS_DICOM-based_Infrastructure

[8] Siddhesh Thakur, et al., "Optimization of deep learning models for inference in low resource environments," ScienceDirect, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0010482525009667

[9] Hassan Sami Adnan et al., "Implementation framework for AI deployment at scale in healthcare systems," ScienceDirect. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004225006674

[10] Ige Gabriel, Adebayo Precious, "Integration of artificial intelligence into clinical workflows: Barriers and best practices," ResearchGate, 2025. [Online]. Available: https://www.researchgate.net/publication/394295429_Integration_of_Artificial_Intelligence_into_Clinical_Workflows_Barriers_and_Best_Practices