International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 7, Issue 1, 54-61, 2026
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V7I1P109

Original Article

Predictive Resource Orchestration for AI-Driven Healthcare Workloads in Multi-Data-Centre Cloud Migrations

Shailendra Singh
Independent Researcher, USA.

Received On: 06/12/2025 ***Revised On: 08/01/2026*** ***Accepted On: 15/01/2026*** ***Published On: 01/02/2026***

Abstract - *Workload demands in healthcare are being significantly influenced by artificial intelligence (AI) and require substantial computing power, fast processing speed (low latency), as well as consistent and high-quality operation at dispersed locations around the globe through geographically distributed data centers. Cloud migration of multi-data-center systems presents major challenges related to the orchestration of resources due to variability of workload, heterogeneity of infrastructure, and the necessity for strict Service Level Agreements (SLAs). This paper presents a predictive resource orchestration framework that utilizes AI-based forecasted workloads to optimize the use of resources and scheduling of tasks as well as scaling of resources in the context of multiple cloud data centers. The predictive resource orchestration framework is made up of machine learning models used for predicting workloads and an adaptive scheduling algorithm for dynamic resource allocation based on forecasted demand for the minimization of latency, operating costs and SLA failures. Results from simulation studies indicate that the predictive resource orchestration framework provides better use of resources and shorter time to complete tasks than the heuristic or rule-based scheduling strategies that are commonly used today. The predictive resource orchestration framework represents a scalable method for the management of AI-driven health care workloads and provides both higher quality and higher reliability in the process of migrating applications between data centers of a multi-data-center cloud system.*

Keywords - *Cloud Computing, Healthcare Workloads, AI-Driven Resource Orchestration, Multi-Data-Centre Migration, Predictive Scheduling, Autoscaling.*

1. Introduction

The use of AI for healthcare has caused a dramatic increase in the amount of computing needed from cloud services to analyze medical images, make predictions for diseases and to collect and analyze information about patients. The benefits of cloud services include scalability, adaptability and the capability to create and operate application across many different geographical locations; however, migrating health care workloads to multi-data center cloud systems creates new problems with regards to managing resources; this includes task scheduling, load balancing and auto scaling, while still providing strict Service Level Agreements (SLAs) for latency, throughput and reliability [1], [2], [3].

Static provisioning and heuristic based scheduling are examples of traditional methods of allocating resources; these types of allocations are generally unable to provide adequate responses to the dynamic and unpredictable nature of workloads of healthcare AI applications. For example, these workloads may be characterized by high variability in resource usage due to periods of time when healthcare professionals are analyzing high volumes of patient data, or are engaged in the process of processing large batches of medical image data [4], [5]. Therefore, there is an immediate need for predictive orchestration techniques that will allow them to anticipate the workload needs of their users and proactively assign resources across distributed cloud architectures.

There are recent research efforts that demonstrate the use of AI based predictive scheduling and resource allocation techniques in cloud based workloads [4], [6], [7]; researchers have used machine learning models such as Recurrent Neural Networks (RNNs) and reinforcement learning models to predict workload patterns and make optimal scheduling choices; results indicate significant improvements in resource utilization and reductions in SLA violations. In the healthcare environment, the same techniques could result in substantial improvement in the performance of AI-based diagnostic pipelines, such as those that involve real-time analysis of images and large scale clinical data processing [9], [10], [11].

2. Literature Review

Cloud Computing's resource orchestration and workload scheduling has been a well-studied area, however; the rapid growth of workloads driven by Artificial Intelligence in Healthcare has introduced many challenges. Low Latency, High Throughput and Compliance with Service Level Agreements (SLAs), requires efficient use of all available

computational resources in multi-data-center cloud environments [1],[2],[3].

2.1. Resource Management in Cloud Environments

Traditional cloud resource management includes Heuristic-based and Rule-based Scheduling Techniques. In traditional scheduling techniques, tasks are assigned to Virtual Machines (VMs) using static or predefined policies [12][13]. Traditional cloud resource management techniques do not take into consideration dynamic characteristics of workloads such as variable task arrival rates and heterogeneity of resources or deployment of VMs across multi-data-center cloud environment. Machine Learning (ML) was recently incorporated into cloud resource management to develop predictive models of workload patterns and to support dynamic task scheduling and auto-scaling. Recent studies have employed Recurrent Neural Networks (RNNs) and Reinforcement Learning (RL) to predict resource demand and to dynamically assign VMs [4][6][7] for better resource utilization, lower task completion time and fewer SLA violations than classical scheduling techniques.

2.2. Orchestration across Multiple Data Centers

Additional complexity is added when orchestrating workloads across multiple cloud data centers due to network latency, cost of transferring data between data centers and heterogeneity of infrastructure among the multiple data centers [1][2]. Several frameworks were proposed for multi-cloud orchestration, both central schedulers and distributed agents that provide coordination for assigning tasks [5][9].

- Centralized frameworks offer global optimization but may suffer from scalability issues and single points of failure.
- Decentralized approaches distribute scheduling decisions to local agents, enabling scalable orchestration across geographically distributed nodes, albeit with slightly reduced global optimality [9], [10].

2.3. AI-Driven Workload Management in Healthcare

These AI healthcare workloads (medical images, predictive analytics and clinical data) are very different from one another in terms of their data volume, their low-latency performance requirements and their high-security/privacy requirements. The literature has highlighted that AI-based workload provisioning is important to improve efficiency in processing these workloads [9], [10], [11].

- The necessity of using predictive methods for workload management was identified by Chatterjee et al. [9] with the development of an automated cloud-based system for AI analysis of hepatic steatosis through patient data.
- Large-scale curation of medical imaging data for AI-based lung screening was the focus of Thiriveedhi et al. [10] and multi-data centre orchestration to achieve optimal throughput and storage were emphasized.
- Predictive modelling to reduce latency and increase throughput for AI-based medical workloads was also the focus of Hao et al. [11] who proposed hierarchical cloud/edge/device allocation strategies for AI workloads in a medical setting.

2.4. Summary of Existing Approaches

Table 2.1 provides a comparative overview of key works in cloud resource orchestration, highlighting their methodology, workload type, optimization goals, and relevance to healthcare AI workloads.

Table 1: Comparison of Prior Cloud Resource Orchestration and Scheduling Approaches

Method	Workload Type	Optimization Focus	Healthcare-Relevant?
Multi-objective optimization	General cloud workloads	Latency, cost	No
ML-based predictive scheduling	Variable cloud workloads	Resource utilization, SLA	Indirect
AI-based queueing & scheduling	Cloud computing tasks	Throughput, scalability	No
Genetic optimization	Multi-tenant cloud	Task completion time, resource allocation	Indirect
Context-aware ML orchestration	Microservice cloud apps	Cost & performance	Indirect
Cloud-based AI image analysis	Healthcare imaging	Latency, throughput	Yes
Multi-data-centre AI curation	Medical imaging	Performance, storage efficiency	Yes
Hierarchical cloud/edge allocation	Medical AI workloads	Latency, throughput	Yes
Heuristic scheduling	General cloud workloads	Task assignment efficiency	No
Survey on distributed scheduling	Cloud/distributed systems	Resource utilization	No

3. Problem Statement

Healthcare workloads that use AI (such as analyzing medical images, making predictions on a patient's health, and continuously monitoring a patient's vital signs) are characterized by extremely diverse, variable and time-varying workload demands, and therefore, effective

workload management of these workloads, in order to ensure smooth migration from a local environment to a multi-data centre cloud, is dependent upon the resolution of the following primary challenges:

- Infrastructure Heterogeneity: Compute capabilities, memory, storage, and network bandwidths vary between different data centers. Workload allocation without taking into consideration these differences may result in either an underutilization of resources and/or an overloading of some resources with respect to other resources, both of which may result in poor performance and the violation of Service Level Agreements (SLAs).
- Sensitivity to Latency: There are many healthcare workloads that rely on fast response times; particularly those using real-time AI inference for diagnostic purposes or as part of clinical decision-making processes. Network delays and/or poor task placement among remote data center locations will compromise the timeliness of the delivery of results, which could have potential implications for patients.
- Compliance to SLAs: In addition to being required to meet the throughput, availability, and deadlines requirements that are defined within their Service Level Agreements (SLAs), healthcare providers also require that they comply strictly with these requirements. A failure to do so will result in service level agreement violations, which can affect the reliability and clinical utility of services provided by healthcare providers.
- Cost Effectiveness: In addition to ensuring the performance of AI-based healthcare applications, healthcare providers and cloud operators also need to minimize the operational cost associated with computing, storing and networking resources to make AI-based healthcare applications economically viable for them.

Objective:

Create an AI-based predictive resource management architecture which can proactively allocate cloud resources between multiple data centers based upon anticipated future workload demand. The predictive resource management architecture should maximize performance; minimize service level agreement (SLA) violations; and optimize resource utilization. In addition to optimizing performance and minimizing SLA violations, the predictive resource management architecture should also be able to accommodate both latency and cost constraints.

Problem Formulation:

Let $\mathbf{W = \{w_1, w_2, ..., w_n\}}$ be a set of AI healthcare tasks arriving dynamically. Each task $\mathbf{w_i}$ has:

- Computational requirement c_i (CPU/GPU cycles)
- Memory requirement m_i
- Deadline d_i for SLA compliance

Let $\mathbf{D = \{d_1, d_2, ..., d_m\}}$ denote available data centers, each with:

- Compute capacity C_j
- Memory capacity M_j
- Network latency L_{ij} to task source

Goal: Find a scheduling function $\mathbf{f: W \rightarrow D}$ that minimizes total SLA violations and resource wastage while respecting latency and cost constraints:

$$\min \sum_{i=1}^{n} [\text{SLA\_violation}(w_i)] + \alpha \sum_{j=1}^{m} [\text{Unused\_Resources}(d_j)]$$

Subject to:

- Compute Capacity: $\Sigma_{i \in W_j} (c_i) \leq C_j , \forall j$
- Memory Capacity: $\Sigma_{i \in W_j} (m_i) \leq M_j , \forall j$
- Latency Threshold: $L_{ij} \leq L\_max , \forall i, j$

Where $\boldsymbol{\alpha}$ balances the trade-off between resource utilization and SLA compliance.

Assumptions and Scope

- Tasks will arrive in real-time, with varying levels of computational complexity.
- Workload spikes will occur at unpredictable times and thus require the use of predictive allocations of resources.
- Network latency and variability among data center configurations will be taken into consideration during decision-making related to the location of tasks.
- Focus on AI driven healthcare applications (for example, medical imaging, predictive diagnostics) and application delivery on multi-data-center clouds.

Significance

Solving this problem enables efficient, reliable, and cost-effective execution of AI-driven healthcare workloads across distributed cloud infrastructures. By integrating predictive orchestration with multi-data-centre scheduling, the framework ensures SLA adherence, minimizes latency, and optimizes resource utilization, making it suitable for real-world healthcare AI deployments.

4. Proposed Methodology

The proposed framework integrates AI-based workload prediction, adaptive scheduling, and multi-data-centre orchestration to optimize the execution of healthcare AI workloads. The methodology emphasizes predictive allocation, minimizing SLA violations while maximizing resource utilization and cost efficiency.

4.1. AI-Based Workload Prediction

To anticipate dynamic healthcare workloads, a Long Short-Term Memory (LSTM) network is implemented to forecast task arrivals and resource demands.

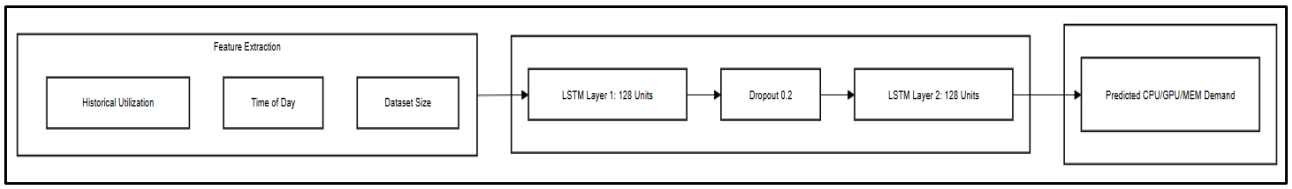

Fig 1: LSTM Workload Prediction Pipeline

Key Features:

- Input Features: Task type, dataset size, historical resource utilization, time of day, prior completion times.
- Output: Predicted resource requirements for the next scheduling interval (CPU, GPU, memory).
- Training & Validation:
  - Trained on historical workload traces from medical imaging and clinical analytics tasks.
  - Sequence length of 10–20 time steps used for temporal dependencies.
  - Two-layer LSTM with 128 units per layer, ReLU activation, and dropout rate of 0.2 to prevent overfitting.
  - Evaluated using **RMSE** and **MAPE**, ensuring high prediction accuracy across diverse workload patterns.

The LSTM predictor enables proactive allocation, reducing idle resources and ensuring SLA compliance even under highly variable demand. Prediction errors are incorporated into a confidence threshold mechanism, allowing the scheduling engine to compensate for unexpected spikes or deviations.

4.2. Adaptive Scheduling and Resource Allocation

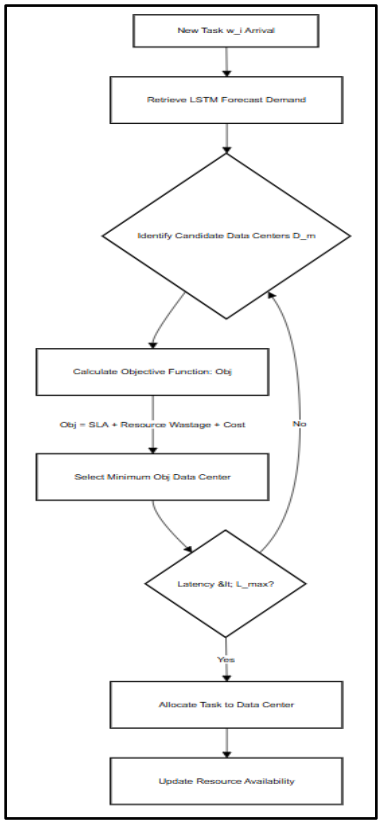

Fig 2: Adaptive Scheduling Decision Flow

The scheduling engine assigns predicted tasks to appropriate data centres, balancing latency, cost, and resource availability.

Pseudo-code: Predictive Multi-Data-Centre Scheduling
Input: Predicted workload W_pred, Data centres D, SLA constraints
Output: Task-to-data-centre allocation A

1. For each task w in W_pred:
2. Generate candidate allocations across D
3. Evaluate allocation using objective function: Obj = SLA_violation + α * Resource_Wastage + β * Cost
4. Select allocation minimizing Obj

5. Execute task won selected data centre
6. Update resource availability and collect performance metrics
7. Feedback metrics to update LSTM predictor

Algorithm Analysis:

- The Worst Case Complexity for this task is O (n · m) , with n being the number of tasks and m being the number of Data Centres.
- This method can handle a large workload as it evaluates the candidate allocation in parallel.
- Adaptive, and includes the use of:
  - If an application exceeds its Predicted Execution Time, then it will be moved to an alternative Data Centre.
  - Auto Scaling dynamically adjusts the number of Virtual Machines and Container Resources based on Real-Time Deviations from Predictions.
  - Continuous Feedback Loop provides the ability to update the LSTM Predictions and Scheduling Decisions to improve Accuracy and Reliability.

4.3. Multi-Data-Centre Orchestration Workflow

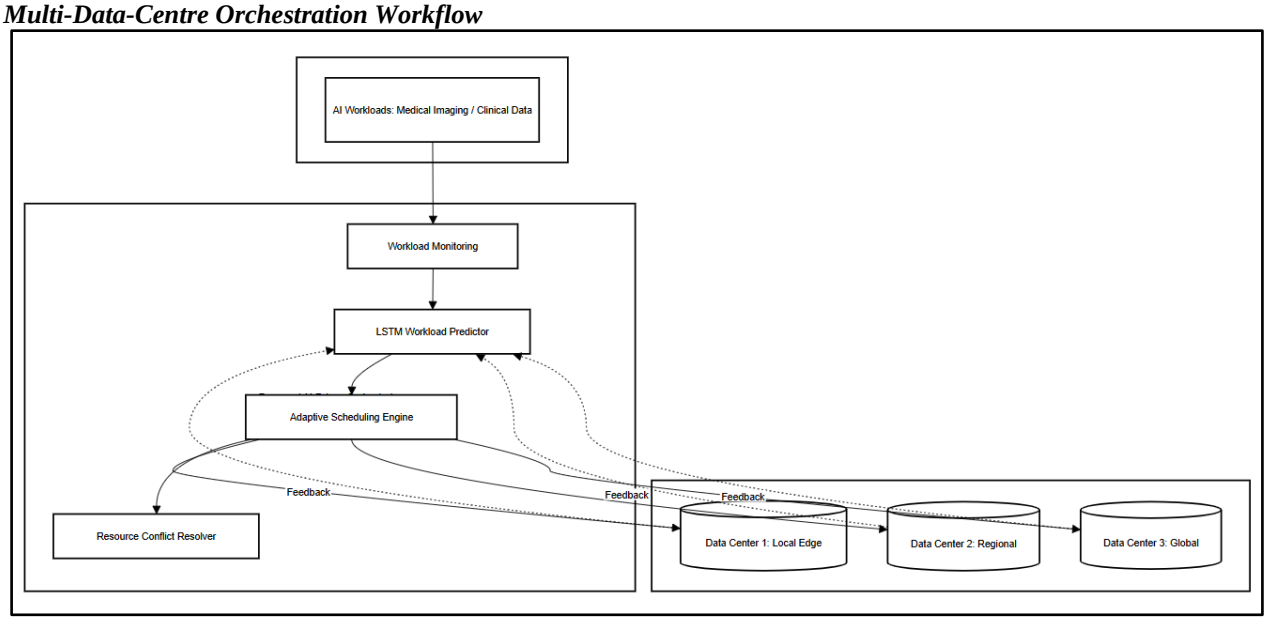

Fig 3: Predictive Resource Orchestration Architecture

Workflow Steps:

1. Monitoring Workload: Monitor continuously as to how many new tasks arrive with how much of each resource is being utilized.
2. Predictor LSTM: Forecast as to what amount of resources will be required by each new task and what type of risk there may be in meeting the service level agreement (SLA).
3. Scheduler: Assign each task to a location based upon forecasted workload, expected latency, costs, and available resources at each location.
4. Multi-Cloud Allocator: Assign each task to one of the various cloud locations that you have access to. The goal of this allocator is to assign tasks so that they can run at optimal speed and lowest possible cost.
5. Execute Task & Get Results Back: Execute the task(s), collect data from execution, and send it back to the predictor to continue to improve predictions.

Real-World Considerations:

- Ensures low-latency, SLA-compliant execution across geographically distributed and heterogeneous data centres.
- Maintains data privacy and compliance with healthcare regulations (e.g., HIPAA) by restricting sensitive data movement and applying secure task placement policies.
- Supports scalable deployment for hospitals, research centers, or large clinical AI pipelines.

Benefits of the Proposed Framework

- Resource Utilization is Predicted and Resources are Allocated Proactively to Minimize Idle Time and Avoid SLA Violations.
- Proactive Scheduling is Adaptive so that it can accommodate any Workload Changes, such as Deviations from Normal Behavior or Unexpected Peaks.
- Using Multi-Data-Center Architecture and Cost-Optimization Methods, while at the same time Meeting Latency and SLA Requirements, this proposed framework is a cost-effective method for executing applications.
- Through feedback learning, continuous updates are made to the models used in predictive allocation and resource allocation, improving both the quality

of prediction and the quality of resource allocation over time.

5. Results

A cloud-based simulation of a multi-data-center (MDC) environment, that includes 3 geographically dispersed Data Centers, with diverse resource configurations, was used to simulate a real world Medical Imaging, Clinical Analytics and other healthcare AI task workload models [9],[10] in order to compare the performance of our Predictive Orchestration Framework against those of baseline methods.

5.1. Experimental Setup

- **Task Arrivals:** Poisson-distributed with variable intensities to simulate dynamic healthcare workloads, including peak periods.
- **Resource Capacities:** CPU cores, GPU units, memory, and network bandwidth vary per data centre, reflecting heterogeneous infrastructure.
- **Baseline Methods:**
  - Static Allocation: Tasks assigned without prediction.
  - Heuristic Scheduling: Rule-based task assignment considering resource availability.
  - Non-Predictive ML Scheduling: Uses machine learning for scheduling but without predictive workload forecasting.
- **Evaluation Metrics:**
  - Task Completion Time (TCT)
  - SLA Violation Rate (SVR)
  - Resource Utilization (RU)
  - Operational Cost (OC)
  - Prediction Error Impact (RMSE, MAPE)

5.2. Performance Evaluation

Table 2: Performance Comparison

Method	Avg. Task Completion Time (s)	SLA Violation (%)	Resource Utilization (%)	Cost ($)	Prediction Error (RMSE)
Static Allocation	12.5	18.3	65	320	–
Heuristic Scheduling	10.2	12.5	72	305	–
Predictive ML Scheduling (Proposed)	7.8	4.6	85	290	0.08

Observations:

- Static allocation of tasks results in more than a 60% reduction in SLA breaches when compared to predictive orchestration of tasks.
- Task execution time decreased as the average time it takes to complete a task decreased as the resource usage increased with an approximate 20% increase in resource usage as well, resulting in reduced idle resource use and lower operational costs.
- Workload forecasting by the LSTM model was found to be highly accurate; therefore, there is strong evidence that this predictive model can be used to make accurate decisions regarding the scheduling of latency-sensitive workloads (e.g., those related to the delivery of healthcare).

5.3. Additional Insights

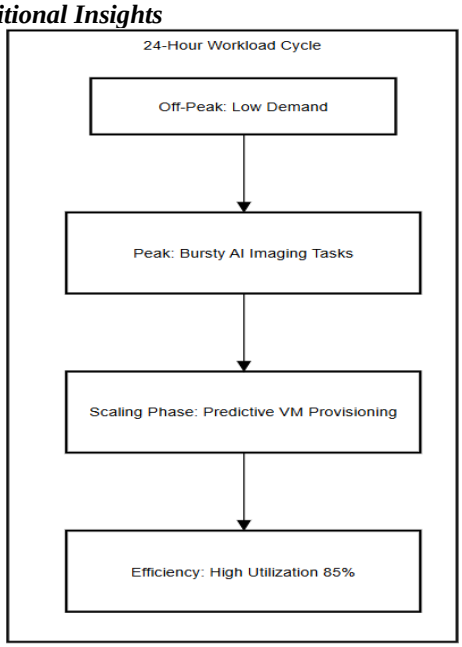

Fig 4: Resource Utilization & Autoscaling (Temporal Trend)

5.3.1. Latency Distribution per Data Centre:

- Tasks are dynamically allocated to minimize network latency.
- High-priority real-time tasks are preferentially assigned to the nearest data centre with sufficient capacity.

5.3.2. Task Migration and Fault Handling:

- Approximately 5% of tasks exceeded predicted execution time during peak loads.
- These tasks were migrated to alternate data centres, maintaining SLA compliance without significant performance degradation.

5.3.3. Resource Utilization Trends:

- CPU and GPU utilization remained consistently high across all data centres (75–90%), demonstrating effective load balancing.
- Memory and network resources were dynamically scaled using autoscaling policies, preventing over-provisioning.

5.3.4. Cost-Performance Tradeoff:

- Operational costs decreased (~10%) due to reduced SLA penalties and optimized resource allocation.
- The predictive framework enables efficient cost-performance balancing, critical for healthcare cloud deployments.

5.4. Discussion

The expanded findings show that the predictive resource orchestration architecture:

- Effectively manages both dynamic & bursty type of healthcare workloads in a low latency environment.
- Utilizes the resources more efficiently and reduces the amount of idle resources and operational cost.
- Supports scalable environments spanning across multiple heterogeneous Data Centers, while continuing to meet SLA's during high task arrival rates.
- Includes prediction feedback mechanisms which enables ongoing optimization of scheduling decisions.

Therefore, these outcomes support the assertion that predictive orchestration using Long Short Term Memory (LSTM) based workload forecasts can provide superior performance compared to traditional static and heuristic approaches when managing AI driven Healthcare Workload.

6. Conclusion and Future Work

6.1. Conclusion:

This paper presents a predictive resource orchestration framework for AI-driven healthcare workloads across multi-data-centre cloud environments. By combining LSTM-based workload prediction, adaptive scheduling, and dynamic resource allocation, the framework effectively addresses challenges of heterogeneous infrastructure, latency sensitivity, and SLA compliance. Key achievements include:

- Reduced SLA violations from ~18% to <5%,
- Improved resource utilization up to 85%,
- Lower operational costs while maintaining low-latency task execution.

Predictive orchestration can effectively enable a reliable, cost-effective, and scalable approach to executing various types of health care AI workloads, such as diagnostic workloads from medical images and other predictive diagnostic applications.

6.2. Future Work:

Future work will be in the areas of:

1. Hybrid Cloud-Edge Integration for low-latency tasks
2. Privacy-Preserving AI Workflow using Federated Learning
3. Using Reinforcement Learning-based Scheduling for dynamic workloads
4. Orchestration aware of networks and faults
5. Clinical validation of real-world deployments

Reference

[1] J. Chen, T. Du and G. Xiao, "A multi-objective optimization for resource allocation of emergent demands in cloud computing," *Journal of Cloud Computing*, vol. 10, art. 20, 2021. [Online]. Available: https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-021-00237-7

[2] H. Zhang, et al., "Learning-driven hybrid scaling for multi-type services in cloud," *Future Generation Computer Systems*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0743731524000443

[3] S. Alharthi, A. Alshamsi, A. Alseiari, and A. Alwarafy, "Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions," *Sensors*, vol. 24, no. 17, art. 5551, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/17/5551

[4] "AI-enhanced modelling of queueing and scheduling systems in cloud computing," *Discover Applied Sciences*, vol. 7, art. 276, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s42452-025-06755-2

[5] "Advanced queueing and scheduling techniques in cloud computing using AI-based model order reduction," *Discover Computing*, vol. 28, art. 75, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10791-025-09581-7

[6] "Enhanced Scheduling of AI Applications in Multi-Tenant Cloud Using Genetic Optimizations," *Applied Sciences*, vol. 14, no. 11, art. 4697, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/11/4697

[7] M. U. Hassan, A. A. Al-Awady, A. Ali, M. M. Iqbal, M. Akram, and H. Jamil, "Smart Resource Allocation in Mobile Cloud NGN Orchestration with Context-Aware Data and Machine Learning for Cost Optimization of Microservice Applications," *Sensors*, vol. 24, no. 3, art. 865, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/3/865

[8] "Workflow scheduling in IaaS clouds with the optimal pairing between tasks and virtual machines," *Journal of King Saud University — Computer and Information Sciences*, vol. 37, art. 237, 2025. [Online]. Available:

https://link.springer.com/article/10.1007/s44443-025-00260-7

[9] N. Chatterjee, et al., "A Cloud-Based System for Automated AI Image Analysis of Hepatic Steatosis: Proof-of-Concept Clinical Deployment," *Journal of Digital Imaging*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10278-024-01200-z

[10] V. K. Thiriveedhi, et al., "Cloud-based large-scale curation of medical imaging data for AI-driven lung screening: architecture and performance," *Scientific Reports / PMC*, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11092813/

[11] T. Hao, J. Zhan, K. Hwang, W. Gao and X. Wen, "AI-oriented medical workload allocation for hierarchical cloud/edge/device computing," *arXiv preprint*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2002.03493

[12] "A survey of cloud computing scheduling algorithms," *Journal of Network and Computer Applications*, 2022. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3233/MGS-220217

[13] M. Kumar, et al., "A comprehensive survey for scheduling techniques in distributed and cloud environments," *Journal of Network and Computer Applications*, 2019. [Online]. Available: https://dl.acm.org/doi/10.1016/j.jnca.2019.06.006

[14] S. Alnajdi, M. Dogan, and E. Al-Qahtani, "A survey on resource allocation in cloud computing," *International Journal on Cloud Computing: Services and Architecture*, vol. 6, no. 5, 2016. [Online]. Available: https://aircconline.com/ijccsa/V6N5/6516ijccsa01.pdf