

Assignment 2

Course Title: Research Methodology in Computer Science

Course Code: AI800

Weighting: 30%

Lecturer: Dr. Mohammed Anbar

Hand-in Date: 19 July 2025

Hand-in Method: Online Submission via E-Learning.

Research Title:

Enhancing Cloud Computing Performance Using Machine Learning Techniques

Problem Statement:

Cloud computing has become a fundamental technology for modern IT systems, offering flexibility, scalability, and cost-effectiveness. However, the rapid growth in user demand and data-intensive applications has exposed performance limitations in traditional optimization approaches. Conventional techniques, such as static resource allocation and rule-based scheduling, are not adaptive enough to handle dynamic workloads and unpredictable traffic. This often results in inefficient resource utilization, increased latency, and higher operational costs.

Recent studies suggest that machine learning (ML) techniques can significantly enhance cloud performance by enabling predictive workload management, intelligent resource allocation, and automated scaling. Approaches like reinforcement learning and workload prediction models have demonstrated better resource efficiency and reduced response times in experimental settings. Despite these advancements, existing solutions are often designed for specific scenarios or evaluated in controlled environments, limiting their practical applicability and scalability in real-world cloud ecosystems.

Therefore, there is a pressing need for research that explores how machine learning techniques can be effectively integrated to improve cloud computing performance, reduce operational costs, and deliver higher quality of service.

Research Objectives:

This research is aimed to investigate the role of machine learning techniques in enhancing the performance of cloud computing systems. To achieve this goal, we have the following objectives:

- Analyze the limitations of traditional cloud optimization techniques and their performance challenges.
- Review and classify machine learning algorithms used for workload prediction, resource allocation and auto-scaling.
- Design a conceptual framework that combines ML-based prediction and optimization strategies for cloud systems.
- Evaluate potential performance improvements, including efficiency and cost reduction, using ML-driven methods.
- Provide recommendations for implementing ML-based solutions in practical, large-scale cloud environments.

Literature Review:

Cloud computing has emerged as a critical component of modern IT infrastructure, enabling organizations to access scalable and on-demand resources. However, as demand for cloud services continues to rise, traditional resource management methods have struggled to maintain efficiency and responsiveness. Several studies have explored how machine learning can address these issues by introducing intelligent, adaptive solutions.

Birhade et al. (2025) presented a theoretical review of AI and ML methods in cloud optimization, emphasizing their potential in dynamic resource provisioning, energy management, and predictive analytics. Although this study provides a broad overview, it lacks experimental validation, highlighting a gap between theoretical models and practical applications.

Nawrocki and Smendowski (2024) conducted a systematic survey of over 70 optimization methods, classifying them into proactive, data-driven, and ML-based approaches. Their analysis highlights how ML methods stand out in handling dynamic workloads and adapting to changing demands.

Sharma and Chaturvedi (2021) examined classical optimization strategies like load balancing and serverless architectures. While their study does not utilize ML, it contextualizes the limitations of traditional methods, further justifying the need for intelligent approaches.

Wang and Yang (2025) developed a hybrid ML model combining LSTM for workload prediction with a Deep Q-Network for resource allocation. Their results show a 32% improvement in resource utilization and a 43% reduction in response time.

Zhou et al. (2021) explored deep reinforcement learning for resource scheduling, achieving significant improvements over traditional scheduling techniques. Similarly, Huang et al. (2023) introduced CloudProphet, a lightweight ML-based tool for performance prediction before deployment.

These studies collectively demonstrate the potential of machine learning in cloud performance optimization, but they also highlight gaps in real-world deployment, scalability, and integration across heterogeneous environments.

References:

- Birhade, A., Shejul, V., & Jadhav, D.R.D. (2025). AI and Machine Learning in Cloud Optimization.
- Nawrocki, P., & Smendowski, M. (2024). A Survey of Cloud Resource Consumption Optimization Methods. *Journal of Grid Computing*, 23(1), p.5.
- Sharma, S., & Chaturvedi, R. (2021). Optimizing Scalability and Performance in Cloud Services: Strategies and Solutions. *ESP Journal of Engineering & Technology Advancements*.
- Wang, L., & Yang, F. (2025). Intelligent Resource Allocation Optimization via ML. *IEEE Transactions on Cloud Computing*.
- Zhou, Y., et al. (2021). DRL-Based Resource Scheduling in Cloud Computing. *Journal of Cloud Computing*.
- Huang, D., Costero, L., Pahlevan, A., Zapater, M., & Atienza, D. (2023). CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds. *IEEE Transactions on Sustainable Computing*, 9(4), 661-676.