

Course Title: Research Methodology in Computer Science

Course Code: AI800

Weighting: 20%

Lecturer: Dr. Mohammed Anbar

Hand-in Date: 12 July 2025

Hand-in Method: Online Submission via E-Learning

Assignment Report

Research Title:

Enhancing Cloud Computing Performance Using Machine Learning Techniques

Student ID: 252001445

Student Name: Abdullah Mohammed Alshahrani

Table of contents:

- Introduction
- Summary and Critical Analysis
- Connections Between the Studies
- Conflicts Between the Findings
- Synthesis Matrix for Literature Review
- References

Introduction:

Cloud computing has become the backbone of modern IT infrastructure which offers scalable, flexible and on-demand computing resources. However, as user demands and data volumes continue to increase and therefore optimizing the performance of cloud services has emerged as a critical challenge. Traditional resource management and scheduling techniques often fall short in dynamic and complex environments which lead to resource underutilization, increased latency and higher operational costs.

In recent years, machine learning (ML) has shown significant potential in enhancing various aspects of cloud performance from intelligent resource allocation and workload prediction to automated scaling and fault detection, ML-based methods offer adaptive and data-driven solutions that outperform static or rule-based approaches.

In the following, we provide literature review that explores how machine learning techniques are applied to improve the performance of cloud computing systems. We consider six peer-reviewed research papers to be reviewed and analyzed in order to highlight current trends, methodologies and findings in this domain. This review aims to provide a comprehensive understanding of the strengths, and limitations of each approach and how these studies collectively contribute to the advancement of cloud optimization strategies.

Summary and Critical Analysis:

Paper 1: AI and Machine Learning in Cloud Optimization (Birhade et al., 2025)

Summary:

This paper provides a conceptual and theoretical review of how artificial intelligence (AI) and machine learning (ML) techniques can be applied to optimize various dimensions of cloud computing performance. Some area has been discussed by authors such as dynamic resource provisioning, predictive analytics for workload management, energy efficiency and cost optimization using AI-driven models. The review positions ML not only as an enhancement tool but as a foundational element in the evolution of adaptive and self-aware cloud systems.

Critical Analysis:

The strength of this paper lies in its breadth and clarity in explaining how ML integrates into different layers of cloud architecture. However, the lack of experimental validation or empirical testing limits its practical applicability. Despite this, the study serves as a solid theoretical foundation, offering a holistic view of the potential directions in ML-driven cloud optimization, and prepares the reader to better understand the implementation-focused studies that follow.

Paper 2: A Survey of Cloud Resource Consumption Optimization Methods (Nawrocki & Smendowski, 2024)

Summary:

A systematic survey of over 70 academic contributions on cloud resource consumption and optimization techniques had been conducted in this paper. The surveyed methods have been classified into three main categories which are: proactive resource management, data-driven optimization strategies and solutions based on machine learning. Each category is analyzed in regarded to its effectiveness, limitations and practical deployment challenges in real-world cloud environments.

Critical Analysis:

It is clear that the paper's comprehensive scope and analytical approach provide valuable context for understanding where ML fits into the broader optimization landscape. However, the absence of implementation case studies or simulation data clearly make the paper largely descriptive. Furthermore, it places a critical foundation by mapping the state of the art and identifying research gaps, specifically in areas where ML can significantly enhance decision-making in cloud performance.

Paper 3: Optimizing Scalability and Performance in Cloud Services (Sharma & Chaturvedi, 2021)

Summary:

This study explores classical and infrastructure-focused approaches to cloud performance improvement, including auto-scaling mechanisms, load balancing techniques, serverless architecture and edge computing. The authors compare these techniques across different deployment environments to assess their impact on scalability, latency and resource efficiency.

Critical Analysis:

Although the paper does not incorporate ML techniques, it contributes significantly by contextualizing the architectural bottlenecks that machine learning models aim to solve. It bridges theoretical performance concerns with practical deployment insights which help to identify scenarios where ML can complement traditional optimization methods. It can be seen that the main limitation is the generality of the discussion which leads to lack in detailed implementation results or focus on adaptive technologies.

Paper 4: Intelligent Resource Allocation Optimization via ML (Wang & Yang, 2025)

Summary:

This paper proposes a hybrid machine learning model combining Long Short-Term Memory (LSTM) for workload prediction and Deep Q-Networks (DQN) for resource allocation decisions. The authors tested the model in a simulated cloud environment which reports improvements such as a 32% increase in resource utilization and a 43% reduction in response time.

Critical Analysis:

This work exemplifies the practical application of ML in cloud computing. It translates the theoretical potential into a functional system that can adapt to real-time changes in resource demand. The integration of prediction and reinforcement learning demonstrates how intelligent models can outperform rule-based systems. However, evaluating the model on limited datasets and in a controlled environment can doubt its scalability and robustness in production settings.

Paper 5: DRL-Based Resource Scheduling in Cloud Computing (Zhou et al., 2021)

Summary:

This study evaluates the use of Deep Reinforcement Learning (DRL) algorithms for scheduling cloud resources. It provides a comparison of DRL performance with conventional scheduling techniques like Round Robin and First-Come-First-Serve. The outcomes show that DRL-based scheduling significantly improves execution efficiency and reduces energy consumption.

Critical Analysis:

It can be seen that the paper provides strong experimental evidence supporting the superiority of ML models in complex scheduling environments. Its technical depth and comparative approach make it a valuable contribution. However, the practical deployment of such DRL systems requires significant computing power and tuning, which may not be feasible for all organizations. Moreover, it reinforces the trend toward adaptive ML-driven resource orchestration.

Paper 6: CloudProphet – ML-Based Performance Prediction for Public Clouds (Huang et al., 2023)

Summary:

CloudProphet is a machine learning tool that is designed to predict the performance of cloud applications before deployment. By using historical workload data, it forecasts metrics like response time and memory usage which helps developers choose optimal resource configurations and scheduling strategies.

Critical Analysis:

This study adds a proactive layer to performance optimization by focusing on pre-deployment decisions. In contrast with other papers that focus on real-time allocation, CloudProphet enhances decision-making during the planning stage. The model's accuracy and lightweight design are strengths. However, the tool application is seen to be limited to web-based applications and lacks evaluation in multi-cloud or heterogeneous environments.

Connections Between the Studies:

- All papers explore optimization of cloud computing performance, either directly through ML or indirectly through architectural improvements.
- Papers 1, 2, 4, 5, and 6 apply machine learning techniques in different ways, from workload prediction to intelligent resource scheduling.
- Paper 1 and 2 provide broad conceptual and survey-based foundations that support the more implementation-focused approaches in papers 4 and 5.
- Paper 3 emphasizes traditional optimization techniques, helping to highlight the innovation introduced by ML approaches in other papers.
- Paper 6 focuses on prediction before deployment, complementing papers 4 and 5, which operate during runtime.
- All studies identify performance efficiency and resource utilization as critical metrics for improvement.

Conflicts Between the Findings:

- Paper 1 and 2 emphasize theoretical potential, while papers 4 and 5 focus on experimental validation.
- Paper 3 does not include machine learning techniques, which may conflict with the adaptive approaches proposed in other papers.
- Paper 6 limits its application to web-based applications, while other papers aim for broader cloud optimization.
- There is a methodological difference between reinforcement learning approaches (Papers 4 and 5) and supervised learning-based predictions (Papers 1, 2, 6).

Synthesis Matrix for Literature Review:

Paper Number	Purpose	Methodology	Key Findings	ML Usage
Paper 1	Review of ML in cloud optimization	Theoretical review	Outlined areas for ML application	Yes
Paper 2	Survey of optimization techniques	Literature survey	Classified optimization strategies	Yes
Paper 3	Evaluate scalability techniques	Comparative analysis	Explored classical methods	No
Paper 4	Hybrid ML model for allocation	Simulated LSTM + DQN	Improved utilization by 32%	Yes
Paper 5	DRL-based scheduling	Experimental comparison	Reduced energy and time	Yes
Paper 6	Performance prediction	ML prediction tool	Proactive planning tool	Yes

References

Birhade, A., Shejul, V. and Jadhav, D.R.D., 2025. AI and Machine Learning in Cloud Optimization.

Nawrocki, P. and Smendowski, M., 2025. A Survey of Cloud Resource Consumption Optimization Methods. *Journal of Grid Computing*, 23(1), p.5.

Sharma, Saloni & Chaturvedi, Ritesh., 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. and Atienza, D., 2024. CloudProphet: a machine learning-based performance prediction for public clouds. *IEEE Transactions on Sustainable Computing*, 9(4), pp.661-676.