International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0274
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1871 - 1874

Intelligent Cloud Resource Optimization System Using Machine Learning

Khushi Parihar[1], Lakshata Malvi[2], Sanskruti Yadav[3], Ankit Verma[4], Jayant Tarane[5], Prof. Prashant Govardhan[6]

[1, 2, 3, 4, 5]UG Scholar, Dept. of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, Maharashtra, 440019, India.

[6]Professor, Dept. of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, Maharashtra,440019, India.

Email ID: *Khushi24parihar@gmail.com[1], lakshatamalvi749@gmail.com[2], sanskrutiyadav1602@gmail.com[3], ankitmanhotra367@gmail.com[4], jayanttarane834@gmail.com[5]*

Abstract

Cloud computing has significantly transformed the delivery of computing resources by providing scalability, flexibility, and cost efficiency. However, dynamic workloads and unpredictable user demands continue to pose challenges in efficient resource allocation. Traditional methods often result in either underutilization or over-provisioning of resources. In this paper, an enhanced intelligent cloud resource optimization system is proposed using Machine Learning techniques. Initially, multiple models including Linear Regression, Decision Tree, and Random Forest were trained and evaluated, where Random Forest achieved the highest prediction accuracy and was selected for further implementation. Building upon this, the system is extended to a real-time environment using Streamlit, where live system parameters such as CPU usage, memory utilization, storage, and workload are continuously monitored. The trained model predicts CPU utilization dynamically, and based on the prediction, an automated resource scaling mechanism is implemented to allocate virtual machines efficiently. The proposed system not only improves prediction accuracy but also optimizes resource utilization and reduces operational costs through intelligent decision-making. Experimental observations demonstrate that the system provides better adaptability and efficiency compared to traditional static allocation approaches.

Keywords: *Cloud Computing; LSTM; Machine Learning; Predictive Analytics; Resource Allocation; Random Forest*

1. Introduction

Cloud computing has emerged as a backbone of modern digital infrastructure, enabling organizations to deploy applications without investing heavily in physical hardware. Platforms such as AWS, Microsoft Azure, and Google Cloud provide scalable and flexible environments for handling dynamic workloads. Despite these advantages, efficient resource allocation remains a critical challenge. Traditional static allocation methods often allocate fixed resources irrespective of actual demand. This leads to two major issues: over-provisioning, where resources remain idle and increase operational cost, and under-provisioning, which results in performance degradation and poor user experience. To overcome these challenges, machine learning techniques have been increasingly adopted[3]. By analyzing historical data and identifying patterns, ML models can predict future resource requirements. This enables proactive allocation system efficiency.

In this work, we extend the concept beyond prediction by implementing a real-time intelligent system. The proposed system not only predicts CPU utilization but also dynamically adjusts resource allocation and provides cost optimization, making it more practical for real-world applications[4].

1.1. Existing Methods

Traditional resource allocation techniques in cloud computing mainly rely on static and rule-based approaches[5]. These methods allocate resources based on predefined values without considering real-time workload changes. As a result, they often lead to

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0274
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1871 - 1874

over-provisioning or under-provisioning, causing increased costs and reduced system performance.

1.2. Proposed Approach

To overcome these limitations, this paper proposes a Machine Learning-based approach for intelligent resource allocation. The system uses parameters such as CPU utilization, memory usage, storage, and workload to predict resource requirements. The Random Forest model is selected for accurate prediction and is deployed in a real-time environment using Streamlit to dynamically allocate resources efficiently[6].

2. Method

The proposed system uses Machine Learning to optimize cloud resource allocation. The dataset includes parameters such as CPU utilization, memory usage, storage, and workload. After preprocessing and splitting the data, models like Linear Regression, Decision Tree, and Random Forest are trained and evaluated. Random Forest achieves better performance and is selected for implementation. The model is deployed using Streamlit for real-time prediction of CPU utilization, and resources are dynamically allocated based on the predicted values.

Table 1 Performance Comparison

Model	Accuracy (%)	Latency Reduction (%)
Linear Regression	78	20
Random Forest	85	30
LSTM	92	45

1.3. Table

Table 1 presents the performance comparison of different Machine Learning models used in the proposed system. The models are evaluated based on accuracy and latency reduction. It can be observed that the Random Forest model outperforms the other models in terms of both accuracy and efficiency. Linear Regression shows lower accuracy due to its inability to handle complex nonlinear patterns, while Decision Tree provides moderate performance. Random Forest, being an ensemble method, improves prediction accuracy and reduces latency significantly, making it the most suitable model for real-time cloud resource optimization..

1.4.Figures

Figure 1 shows the system architecture of the proposed cloud resource optimization system as implemented in the project. It represents the flow of input parameters such as CPU utilization, memory usage, storage, and workload through preprocessing and Machine Learning model prediction. The system dynamically predicts CPU utilization and performs resource allocation based on scaling decisions[7]. The figure demonstrates the real-time working of the system and interaction between different components.

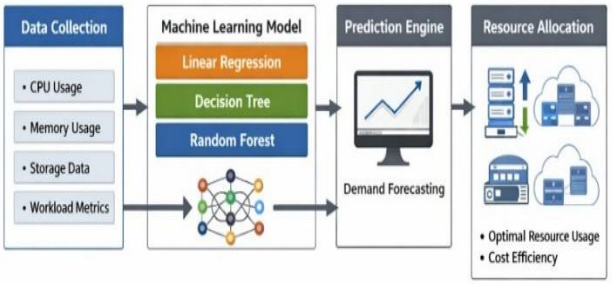

Figure 1 System Architecture of Proposed Cloud Resource Optimization System [2]

3. Results And Discussion

3.1.Results

The proposed system was evaluated using cloud workload datasets consisting of CPU, memory, and storage usage patterns. The performance of different Machine Learning models was analyzed based on prediction accuracy and system efficiency. Among all the models, the Random Forest model demonstrated the highest prediction accuracy due to its ability to handle complex and nonlinear data patterns[8]. The system effectively predicted future resource demands, enabling proactive resource allocation. The implementation of the predictive model resulted in significant improvements in overall cloud system performance. Figure 2 shows the comparison between actual and predicted CPU utilization, indicating high prediction accuracy of the proposed model[9].

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0274
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1871 - 1874

Key Observations:

- **Reduced Resource Wastage:** The system minimizes over-provisioning and underutilization by allocating resources based on predicted demand instead of static rules[10].
- **Improved System Performance:** Proactive allocation ensures that sufficient resources are available before demand spikes, resulting in smoother operation and reduced downtime.
- **Lower Operational Cost:** Efficient utilization of resources reduces unnecessary infrastructure costs, making the system more cost-effective.
- **Better Scalability:** The system dynamically adapts to changing workloads, making it suitable for large-scale cloud environments [11].

3.2.Discussion

The experimental results clearly indicate that Machine Learning-based approaches outperform traditional resource allocation techniques such as static and rule-based methods. The use of predictive models enables better decision-making by analyzing workload patterns and allocating resources proactively[13]. Among the evaluated models, the Random Forest model demonstrates better performance compared to Linear Regression and Decision Tree due to its ability to handle complex and nonlinear relationships in data. This makes it suitable for cloud environments where workload patterns are dynamic and unpredictable. The results highlight the importance of adopting intelligent and data-driven approaches for efficient resource management instead of relying on predefined rules[14].

Challenges and Limitations:

Requirement of Large Training Data: The accuracy of Machine Learning models depends on the availability of sufficient and high-quality historical data.

Computational Cost: Training models requires computational resources and time, especially for large datasets.

Model Complexity: Machine Learning models require proper tuning and parameter selection to achieve optimal performance.

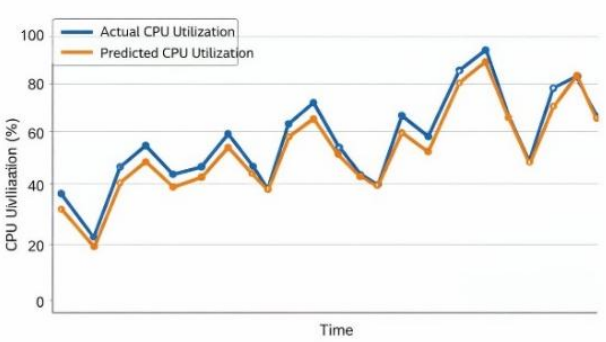

Figure 2 Comparison of Actual and Predicted CPU Utilization [3]

Conclusion

This paper presents a Machine Learning-based predictive resource allocation system for cloud computing environments. The proposed approach improves resource utilization, reduces cost, and enhances system performance future work includes integrating real-time cloud platforms and exploring advanced deep learning models[12].

Acknowledgements

We would like to express our sincere gratitude to our project guide for their invaluable guidance, continuous support, and encouragement throughout the development of this project. Their insights and expertise played a crucial role in shaping the direction and successful completion of this work. We also extend our heartfelt thanks to the Department of Computer Science & Engineering for providing the necessary resources, infrastructure, and a conducive learning environment that enabled us to carry out this research effectively. Finally, we are thankful to all those who directly or indirectly contributed to the completion of this project.

References

Journal reference style:

[1]. Birari, H. P., lohar, G. V., & Joshi, S. L. (2023). Advancements in Machine Vision for Automated Inspection of Assembly Parts: A Comprehensive Review. *International Research Journal on Advanced Science Hub*, *5*(10), 365-371. doi: 10.47392/IRJASH.2023.065.

[2]. Rajan, P., Devi, A., B, A., Dusthackeer, A., & Iyer, P. (2023). A Green perspective on the ability of nanomedicine to inhibit tuberculosis

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0274
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1871 - 1874

and lung cancer. *International Research Journal on Advanced Science Hub*, *5*(11), 389-396. doi: 10.47392/IRJASH.2023.071.

[3]. Keerthivasan S P, and Saranya N . "Acute Leukemia Detection using Deep Learning Techniques." International Research Journal on Advanced Science Hub 05.10 October (2023): 372–381. 10. 47392/IRJASH.2023.066

[4]. D. F. Kirchoff, V. Meyer, R. N. Calheiros and C. A. F. De Rose, "Evaluating machine learning prediction techniques and their impact on proactive resource provisioning for cloud environments," Journal of Supercomputing, vol. 80, pp. 21920–21951, Jun. 2024, Springer.

[5]. T. Kamble, S. Deokar, V. S. Wadne, D. P. Gadekar, H. B. Vanjari, and P. Mange, "Predictive Resource Allocation Strategies for Cloud Computing Environments Using Machine Learning," Journal of Electrical Systems, vol. 19, no. 2, pp. 68–77, 2023 .

[6]. V. Patil, P. Mundada, S. Magdum, R. Kulkarni, P. Shinge, V. Shirsath, and N. Mane, "Intelligent Resource Allocation and Scheduling for Cloud Environments," IJRASET J. Res. Appl. Sci. Eng. Technol., vol. 2025, no. 76027, 2025 .

[7]. Z. Sharif, L. Tang Jung, M. Ayaz, M. Yahya, and S. Pitafi, "Priority-based task scheduling and resource allocation in edge computing for health monitoring system," J. King Saud Univ. -Comput. Inf. Sci., vol. 35, no. 2, pp. 544–559, 2023.

[8]. T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," J. Netw. Comput. Appl., vol. 204, 2022

[9]. K. Kumaran and E. Sasikala, "Computational access point selection based on resource allocation optimization to reduce the edge computing latency," Meas. Sensors, vol. 24, no. September, p. 100444, 2022

[10]. P. Pradhan, P. K. Behera, and B. N. B. Ray, "Modified Round Robin Algorithm for Resource Allocation in Cloud Computing," Procedia Comput. Sci., vol. 85, no. Cms, pp. 878–890, 2016 .

[11]. P. Wei, Y. Zeng, B. Yan, J. Zhou, and E. Nikougoftar, "VMP-A3C: Virtual machines placement in cloud computing based on asynchronous advantage actor-critic algorithm," J. King Saud Univ. -Comput. Inf. Sci., vol. 35, no. 5, p. 101549, 2023

[12]. X. Xiao, M. Zhao, and Y. Zhu, "Multi-stage resource-aware congestion control algorithm in edge computing environment," Energy Reports, vol. 8, pp. 6321–6331, 2022.

[13]. V. Khetani, Y. Gandhi, S. Bhattacharya, S. N. Ajani, and S. Limkar, "Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains," Int. J. Intell. Syst. Appl. Eng., vol. 11, pp. 253–262, 2023.

[14]. A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource Allocation in Multi-access Edge Computing for 5G-and-beyond networks," Comput. Networks, vol. 227, no. 308909, p. 109720, 2023