International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Resource Allocation Using Machine Learning

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

[2]Professor, CSE , Priyadarshini College of Engineering Nagpur, Maharashtra, India
[13456]UG Student, CSE , Priyadarshini College of Engineering Nagpur, Maharashtra, India

---

Abstract -*With the increasing demand for cloud services, optimizing resource utilization is paramount to minimize costs and improve performance. Traditional resource allocation methods often struggle with dynamic workloads and fail to fully leverage available resources. This work explores the application of various machine learning models to predict optimal resource configurations based on diverse workload characteristics. We analyze a dataset containing features such as CPU usage, memory usage, network usage. Several classification algorithms, including Random Forest Classifier, XGBoost Classifier, and Logistic Regression, are implemented and evaluated to determine their effectiveness in predicting the 'Optimized Resource Allocation'. The results demonstrate that machine learning models can effectively learn complex relationships within the data and provide accurate predictions for resource allocation, offering a data-driven approach to improve cloud infrastructure efficiency. This research highlights the potential of machine learning to automate and optimize resource management in cloud environments, leading to significant improvements in performance and cost savings. The comparison of different models provides insights into their suitability for this specific task, guiding future efforts in developing intelligent resource allocation systems. The findings contribute to the growing body of work on applying artificial intelligence to solve real-world cloud computing challenges*.

Key Words: Resource allocation, algorithms, Cloud computing, Machine learning

1. INTRODUCTION

The rapid evolution of cloud computing platforms such as AWS, Azure, and GCP has exposed the limitations of traditional static resource provisioning, which often leads to over-provisioning or under-provisioning and increased operational costs. The growing heterogeneity of cloud workloads further complicates efficient resource management, necessitating adaptive and data-driven approaches. Machine learning (ML) offers an effective solution by leveraging historical workload and performance data to predict future resource demands.

This research formulates optimal cloud resource provisioning as a supervised ML classification problem, where workloads are categorized into predefined optimized resource allocation levels based on CPU, memory, and network usage. Three classification algorithms—Random Forest, XGBoost, and Logistic Regression—are implemented and comparatively evaluated. The methodology includes data preprocessing, model training, and performance assessment using accuracy and classification reports. The study aims to enhance resource utilization and cost efficiency through intelligent ML-based provisioning strategies.

2. Methodology of Review

This review is based on academic research published between 2016 and 2025, sourced from reputed digital libraries such as IEEE Xplore, SpringerLink, Google Scholar, ResearchGate, and ScienceDirect. The studies were selected based on three primary criteria: they must focus on cloud resource management or provisioning, apply machine learning or artificial intelligence techniques, and evaluate performance using measurable metrics such as resource utilization, cost efficiency, or system throughput.

A broad range of research articles were analyzed to understand the evolution of intelligent and data-driven cloud resource allocation strategies. Among these, several studies were identified as highly relevant due to their use of supervised learning models, reinforcement learning, and hybrid optimization techniques for dynamic workload management. These selected works were critically examined to compare algorithmic approaches, system architectures, evaluation methodologies, and performance outcomes. The review also highlights existing challenges, including scalability, model generalization, and real-time adaptability, which motivate the need for more efficient ML-based resource provisioning solutions.

3. Literature Review

3.1 Evaluating Machine Learning Prediction Systems and their Impact on Proactive Resource Provisioning for Cloud Environments

Kirchoff et al. (2024) evaluated machine learning prediction techniques to optimize proactive resource provisioning in cloud environments. Their study found that ensemble methods like Random Forest and XGBoost provided higher accuracy and reliability compared to linear models. While effective in reducing resource wastage and improving system performance, the approach relied heavily on historical workload data, which may not capture sudden spikes.

3.2 Predictive Resource Allocation Using Machine Learning

Kamble et al. (2023) proposed predictive resource allocation strategies using supervised learning classifiers for cloud computing. Their framework categorized incoming workloads into predefined resource allocation levels, enhancing system efficiency. However, the approach required extensive feature preprocessing and was primarily validated through simulations.

3.3 Intelligent Resource Allocation and Scheduling for Cloud Environments

Patil et al. (2025) introduced an intelligent framework combining predictive modeling and scheduling to allocate cloud resources adaptively. Their system reduced resource contention and improved responsiveness under dynamic workloads, though multi-cloud scalability and cost optimization were not fully addressed.

3.4 Priority-Based Task Scheduling: Resource Allocation in Edge Health Monitoring

Sharif et al. (2023) developed a priority-based task scheduling and resource allocation method tailored for edge computing in health monitoring systems. By considering task priority and system resource constraints, the framework improved response times and resource utilization, which is critical for real-time patient monitoring. While the method proved effective in experimental setups, it faced challenges under highly variable workloads and heterogeneous device environments, highlighting the need for adaptive ML-based allocation strategies.

3.5 ML-Centric Resource Management: Review and Future Directions in Cloud Computing

Khan et al. (2022) provided a comprehensive review of ML-centric resource management in cloud computing, including supervised, unsupervised, and reinforcement learning techniques. The paper summarized the state-of-the-art in predictive resource allocation, highlighting emerging trends, challenges, and future directions. It emphasized the benefits of using ML for proactive workload prediction and adaptive scaling. However, the review identified a lack of real-world experimental validations, with most studies being confined to simulations or controlled environments.

3.6 Computational Access Point Selection: Reducing Latency via Resource Optimization

Kumaran and Sasikala (2022) investigated computational access point selection to reduce latency in edge computing, emphasizing optimized resource allocation strategies. Their approach aimed to minimize task completion time and improve overall system responsiveness. Results showed significant latency reduction for edge devices. Nevertheless, the method did not integrate predictive ML models for dynamic workload adaptation, limiting its effectiveness under unpredictable load variations.

3.7 Modified Round Robin Algorithm: Efficient Resource Allocation in Cloud Computing

Pradhan et al. (2016) proposed a modified Round Robin algorithm for cloud resource allocation. The approach provided a simple and computationally efficient scheduling mechanism suitable for homogeneous workloads. It improved fairness in task execution order but lacked adaptability to dynamic workloads and did not incorporate predictive insights, which limited its applicability in modern cloud environments with varying demand patterns.

3.8 VMP-A3C for VM Placement: Reinforcement Learning in Cloud Computing

Wei et al. (2023) developed VMP-A3C, an asynchronous advantage actor-critic reinforcement learning model for virtual machine placement in cloud computing. By learning optimal placement strategies from environmental feedback, the method improved resource utilization and task scheduling under dynamic conditions. It outperformed traditional heuristic approaches in simulation studies. However, the model required substantial computational resources for training, which could limit its deployment in large-scale cloud systems.

3.9 Multi-Stage Resource-Aware Congestion Control: Edge Computing Environments

Xiao et al. (2022) designed a multi-stage resource-aware congestion control algorithm for edge computing environments. The approach aimed to maintain system stability under high traffic loads by allocating resources efficiently and preventing congestion. Experimental evaluation demonstrated improved throughput and reduced packet loss. Nonetheless, the algorithm did not incorporate predictive ML-based allocation, limiting adaptability to rapidly changing workloads.

3.10 Cross-Domain ML and DL Analysis: Impact on Diverse Computational Domains

Khetani et al. (2023) conducted a cross-domain analysis of machine learning and deep learning, evaluating their impact across diverse computational fields including cloud resource optimization. The study highlighted how models trained in one domain could potentially be applied in others, demonstrating versatility. While offering valuable insights into the applicability of ML and DL for cloud systems, the study lacked detailed experimental

evaluation in real-time cloud environments, limiting its practical implications.

3.11 Resource Allocation in Multi-Access Edge Computing: 5G and Beyond Networks

Sarah et al. (2023) addressed resource allocation in multi-access edge computing for 5G and beyond networks. Their framework dynamically allocated resources to meet latency and throughput requirements, which is critical for high-bandwidth, low-latency applications. Results showed improved system performance and user experience. However, the method relied on static allocation thresholds and would benefit from integration with predictive ML techniques for more proactive adaptation.

3.12 Efficient Resource Management: Cloud Environment Optimization Techniques

Swain et al. (2022) proposed an efficient resource management framework for cloud environments, focusing on improving execution time and overall resource utilization. Their method incorporated task prioritization and basic scheduling policies to optimize performance. While effective for moderate workloads, the framework relied on static allocation strategies and lacked predictive capabilities, reducing adaptability to sudden workload spikes or highly dynamic environments.

3.13 Priority-Based Resource Allocation: Hybrid Tree-Enhanced Auto-Scaling Framework

Nagamani and Kailasam (2024) developed a hybrid tree-enhanced vector machine model for proactive auto-scaling and workload prediction in cloud computing. Their system combined predictive modeling with priority-based resource allocation, resulting in improved allocation efficiency and reduced response times. The hybrid approach, however, introduced higher computational complexity, which could impact performance in large-scale or resource-constrained systems.

3.14 Optimizing Cloud Resource Allocation: Strategies for Efficient Computing

Anbarkhan (2024) explored machine learning-based strategies for optimizing cloud resource allocation. The framework dynamically adjusted computational resources based on predicted workload patterns, improving overall performance and energy efficiency. While effective, the approach depended on accurate workload prediction models, which may not generalize across heterogeneous cloud environments with unpredictable workloads.

3.15 Intelligent Multi-Agent Reinforcement Learning: Cloud Resource Allocation

Belgacem et al. (2022) presented a multi-agent reinforcement learning model for cloud resource allocation, enabling autonomous learning of optimal strategies in dynamic environments. The system demonstrated scalability and adaptability, improving resource utilization under fluctuating workloads. Nonetheless, implementation complexity and the need for substantial computational resources posed challenges for deployment in heterogeneous, large-scale cloud infrastructures.

4. Comparative Discussion

Recent studies on machine learning-based cloud resource allocation reveal that different models offer distinct advantages and limitations. Ensemble methods like Random Forest and XGBoost consistently achieve high accuracy and robustness, effectively capturing non-linear workload patterns and minimizing resource wastage. Regression models and LSTM networks are effective for workload prediction trends, but they often struggle with abrupt workload spikes and real-time adaptability. Reinforcement learning approaches, on the other hand, provide dynamic and autonomous resource allocation, adapting to changing workloads and maintaining QoS, though they require higher computational resources and longer training times.

Across these studies, common challenges include scalability in large multi-cloud environments, reliance on offline historical data for training, and limited consideration of cost and energy efficiency. The comparative analysis suggests that hybrid approaches—combining ensemble models for accurate prediction with reinforcement learning for adaptive decision-making—could offer a balanced solution. Additionally, incorporating multi-cloud and edge-cloud scenarios, along with SLA and energy considerations, is crucial for practical, large-scale implementations. These insights provide a foundation for the current study, which uses supervised learning classifiers to categorize workloads and optimize resource allocation efficiently.

4. CONCLUSIONS

The research demonstrates that machine learning techniques can play a significant role in optimizing resource allocation within cloud computing environments by effectively learning patterns from historical workload data. By analyzing critical metrics such as CPU utilization, memory consumption, and network usage, the proposed models enable informed and proactive provisioning decisions. Among the evaluated classifiers, ensemble-based approaches—particularly Random Forest and XGBoost—consistently achieved superior accuracy, robustness, and reliability compared to linear models, highlighting their ability to capture complex and non-linear workload behaviours.

The experimental results confirm that data-driven resource allocation significantly reduces resource

wastage, mitigates performance degradation, and enhances overall system efficiency. Furthermore, the comparative analysis validates the suitability of supervised learning techniques for categorizing workloads into optimized resource allocation levels in dynamic cloud environments. This study provides a strong foundation for future research, which may explore real-time implementation, integration with reinforcement learning for adaptive decision-making, and scalability across large-scale multi-cloud platforms. Ultimately, the findings contribute toward the advancement of intelligent, automated, and cost-effective cloud resource management solutions.

5. REFERENCES

[1] 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.

[2] 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 . https://doi.org/10.52783/jes.692

[3] 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 . https://doi.org/10.22214/ijraset.2025.76027

[4] 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, doi: 10.1016/j.jksuci.2023.01.001.

[5] 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

[6] 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

[7] 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 . https://doi.org/10.1016/j.procs.2016.05.278

[8] 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, doi: 10.1016/j.jksuci.2023.04.002.

[9] 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, doi: 10.1016/j.egyr.2022.04.078.

[10] 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.

[11] 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

[12] S. R. Swain, A. K. Singh, and C. N. Lee, "Efficient Resource Management in Cloud Environment," pp. 1–9, 2022

[13] P. Nagamani and S. Kailasam, "Effective priority-based resource allocation for proactive auto-scaling framework in workload prediction using hybrid tree-enhanced vector machine model," *Discover Sustainability*, vol. 5, art. no. 391, 2024 . https://doi.org/10.1007/s43621-024-00583-x

[14] S. H. Anbarkhan, "Optimizing cloud resource allocation with machine learning: Strategies for efficient computing," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 30, no. 1, pp. 1–12, 2024

[15] A. Belgacem, S. Mahmoudi, and M. Kihl, "Intelligent multi-agent reinforcement learning model for resources allocation in cloud computing," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 6, pp. 2391-2404, 2022 .