Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

Energy-Efficient Resource Allocation in Cloud Computing Environments Using Machine Learning

Dr. Suhas Khot, Principal, K. J. College of Engineering and Management Research, Kondhwa, Pune

Omkar Bibhishan Bhalekar, Assistant Professor, K. J. College of Engineering and Management Research, Kondhwa, Pune

Dr. Nikita Janmejay Kulkarni, Associate professor, K. J. College of Engineering and Management Research, Kondhwa, Pune

Akansha Shivaji Kamthe, Assistant Professor, K. J. College of Engineering and Management Research, Kondhwa, Pune

Abstract

Cloud computing has become a fundamental technology for delivering scalable and on-demand computing resources to diverse applications. However, the rapid growth of cloud data centers has significantly increased energy consumption, operational costs, and environmental impact. Efficient resource allocation is therefore essential to improve server utilization while maintaining Quality of Service (QoS). Traditional scheduling algorithms often rely on static or heuristic approaches that are unable to adapt effectively to dynamic workload variations, resulting in resource underutilization and unnecessary power consumption. This paper proposes an Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework that employs supervised machine learning to predict workload demand and perform proactive virtual machine allocation. The proposed framework integrates workload monitoring, feature extraction, prediction, and dynamic resource allocation to optimize energy efficiency while reducing Service Level Agreement (SLA) violations. Performance evaluation is conducted using CloudSim Plus with representative cloud workload traces. Simulation results indicate that the proposed framework reduces energy consumption, improves CPU utilization, minimizes VM migrations, and enhances response time compared with conventional scheduling approaches. The proposed method offers a scalable and sustainable solution for intelligent cloud resource management.

Keywords: Cloud Computing, Resource Allocation, Machine Learning, Energy Efficiency, Virtual Machine Consolidation, and Green Computing.

1. Introduction

Cloud computing has revolutionized the delivery of computing services by providing on-demand access to shared resources such as processing power, storage, networking, and software applications through the Internet. Its scalability, flexibility, and pay-as-you-go pricing model have led to widespread adoption across industries including healthcare, finance, education, manufacturing, and e-commerce. Virtualization technologies enable multiple virtual machines (VMs) to execute on the same physical server, improving hardware utilization and simplifying infrastructure management. The continuous expansion of cloud services has resulted in the

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

construction of large-scale data centers containing thousands of physical servers. Although these infrastructures provide substantial computational capacity, they also consume considerable electrical energy. Studies have shown that even under low utilization, servers continue to draw a significant proportion of their peak power, increasing operational expenditure and contributing to carbon emissions. Consequently, improving the energy efficiency of cloud data centers has become a major objective for cloud service providers and researchers.

Resource allocation is one of the most critical functions in cloud computing because it determines how computational resources are assigned to virtual machines and user applications. Effective allocation should maximize resource utilization while minimizing energy consumption, response time, and SLA violations. Conventional scheduling algorithms such as First-Come-First-Serve (FCFS), Round Robin (RR), and static threshold-based methods are computationally efficient but generally make decisions based on the current system state. As a result, they often fail to respond effectively to rapidly changing workloads, leading to inefficient resource utilization and unnecessary VM migrations. Machine learning has emerged as a promising approach for intelligent resource management by enabling cloud systems to learn workload patterns from historical data and predict future resource requirements. Supervised learning algorithms such as Random Forest, Decision Trees, Support Vector Machines (SVM), and Gradient Boosting have demonstrated strong predictive capability while maintaining relatively low computational complexity. By anticipating workload changes before they occur, predictive resource allocation enables proactive VM placement, improved server consolidation, and reduced energy consumption. In this paper, an Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework is proposed for dynamic cloud environments. The framework combines workload prediction with intelligent VM allocation to improve resource utilization and reduce energy consumption without compromising Quality of Service. The proposed methodology is evaluated through simulation using CloudSim Plus, and its performance is analyzed using energy consumption, CPU utilization, response time, throughput, VM migrations, and SLA violations as evaluation metrics.

2. Review of Literature

Cloud computing has become a dominant paradigm for delivering scalable and on-demand computing services. However, the increasing number of cloud data centers has resulted in high energy consumption and operational costs, making energy-efficient resource allocation a major research challenge. Conventional scheduling algorithms are often unable to cope with dynamic workload variations, motivating researchers to investigate intelligent resource management approaches based on machine learning and artificial intelligence. Beloglazov and Buyya pioneered energy-aware resource allocation through dynamic Virtual Machine (VM) consolidation, demonstrating that workload migration from underutilized servers to fewer physical hosts can significantly reduce power consumption while maintaining Quality of Service (QoS). Their work established VM consolidation as one of the foundational approaches for green cloud computing,

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

although frequent VM migrations may introduce additional overhead and SLA violations under rapidly changing workloads [3,4,5].

Recent advances have shifted from static scheduling toward predictive resource management using machine learning. Khan *et al.* presented a comprehensive review of machine learning-centric resource management in cloud computing and classified existing approaches into workload prediction, VM placement, resource provisioning, VM consolidation, and thermal management. Their survey concluded that supervised machine learning techniques improve resource allocation by learning workload patterns from historical data, but challenges remain in reducing model training time, computational complexity, and scalability for large cloud infrastructures [7].

Similarly, Saxena and Singh reviewed workload forecasting and predictive resource management techniques for cloud environments. They emphasized that accurate workload prediction enables proactive resource allocation, minimizes over-provisioning and under-provisioning, reduces SLA violations, and improves overall energy efficiency. Their study also highlighted that heterogeneous workloads and rapidly fluctuating resource demands remain significant challenges for existing prediction models [8]. Machine learning has been increasingly applied to optimize cloud resource allocation, workflow scheduling, and live VM migration. Singh *et al.* categorized machine learning applications into supervised, unsupervised, and reinforcement learning approaches for cloud computing. Their review reported that predictive models significantly improve dynamic resource allocation, energy optimization, and workflow scheduling compared with traditional heuristic methods. However, balancing energy efficiency with Quality of Service remains an open research problem [9,11]. Deep learning techniques have further enhanced workload prediction accuracy. Recent studies have shown that Long Short-Term Memory (LSTM), convolutional neural networks (CNNs), and reinforcement learning improve adaptive scheduling and feature extraction for dynamic cloud environments. Although deep learning provides superior prediction capability, it generally requires larger training datasets and higher computational resources than conventional machine learning algorithms [13,15].

Energy management has also been investigated using clustering, optimization, and machine learning techniques. Masdari and Khoshnevis proposed a multi-objective optimization policy for cloud resource allocation that balances energy efficiency and system performance. Their findings indicate that integrating optimization algorithms with predictive machine learning techniques improves overall cloud performance. However, many existing approaches primarily focus on reducing energy consumption without simultaneously optimizing response time, throughput, and SLA compliance [14]. Recent studies have confirmed the growing importance of artificial intelligence in cloud computing. AI-driven resource management enables intelligent workload forecasting, dynamic resource allocation, and adaptive scheduling. Predictive allocation mechanisms consistently outperform reactive scheduling because they allocate resources before workload peaks occur, thereby reducing resource wastage and improving service reliability [7,11,13].

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute
InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

Despite these advancements, several research gaps remain. Most existing methods optimize a single performance objective, such as energy consumption or execution time, without simultaneously considering CPU utilization, response time, throughput, VM migration overhead, and SLA violations. Furthermore, deep learning and reinforcement learning approaches often introduce high computational complexity, making them less suitable for real-time cloud scheduling. These limitations motivate the development of the proposed Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework, which employs a lightweight supervised learning model for predictive workload estimation and dynamic VM allocation to improve both energy efficiency and Quality of Service.

3. Research Methodology

3.1 Proposed Framework

The proposed Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework is designed to improve energy efficiency in cloud computing by integrating workload prediction with intelligent resource allocation. Unlike conventional scheduling algorithms that allocate resources based only on the current system state, the proposed framework predicts future workload demands and performs proactive virtual machine (VM) allocation. This predictive strategy reduces unnecessary VM migrations, improves resource utilization, and minimizes energy consumption while maintaining Quality of Service (QoS). The framework consists of five functional modules: workload monitoring, data preprocessing, machine learning prediction, resource allocation, and performance evaluation. Initially, workload information such as CPU utilization, memory usage, storage consumption, and network bandwidth is collected from physical hosts and virtual machines. The collected data are preprocessed to remove inconsistencies, normalize feature values, and prepare the dataset for model training. A supervised machine learning model is then trained using historical workload records to predict future resource demand. Finally, the system continuously evaluates allocation performance using energy consumption, CPU utilization, throughput, response time, VM migration count, and SLA violations.

3.2 Workflow of the Proposed Framework

The workflow of the proposed framework consists of the following sequential stages:

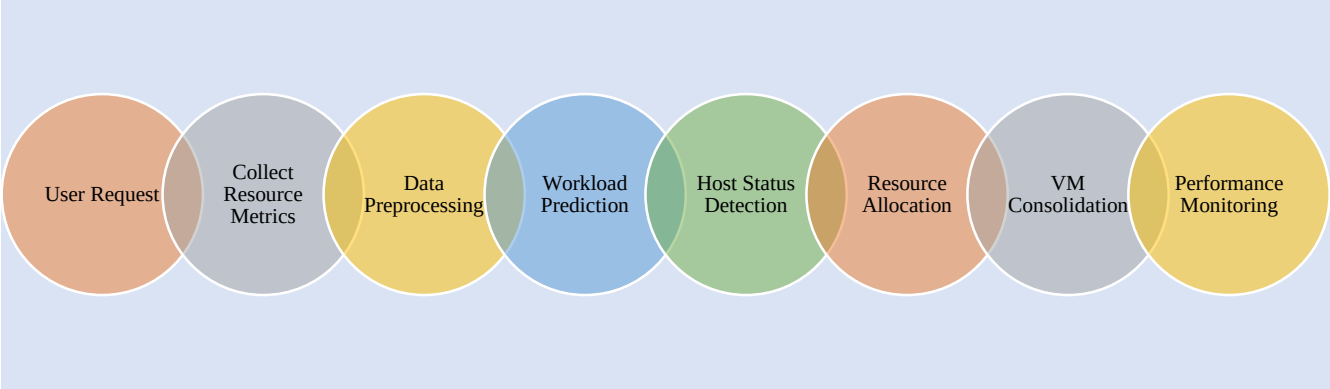

Figure 1. Workflow of the Proposed Framework

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

This closed-loop process enables proactive scheduling decisions and continuous adaptation to changing workload conditions.

3.3 Machine Learning Model

The predictive component of the proposed framework employs a Random Forest Regressor because of its robustness, ability to model nonlinear relationships, and relatively low computational complexity. Random Forest constructs multiple decision trees using bootstrap sampling and combines their predictions to estimate future workload demand. Compared with individual decision trees, ensemble learning reduces overfitting and improves prediction stability.
The input feature vector is represented as

$$X = \{CPU, Memory, Storage, Bandwidth, ActiveVMs\}$$

where each feature corresponds to the current utilization of a cloud host.
The prediction model estimates future resource demand as

$$\hat{Y} = f(X)$$

where $f(\cdot)$ denotes the trained Random Forest model and $\hat{Y}$ represents the predicted workload for the next scheduling interval.

3.4 Mathematical Model

The allocation problem is modeled using a set of physical hosts

$$H = \{H_1, H_2, \dots, H_n\}$$

and a set of virtual machines

$$V = \{V_1, V_2, \dots, V_m\}.$$

Each virtual machine is assigned to a physical host through the allocation function

$$A(V_i) = H_j.$$

The CPU utilization of a host is calculated as

$$U_{cpu} = \frac{\sum_{i=1}^{m} C\,PU_i}{CPU_{total}}$$

where CPU_i denotes the CPU demand of the i^{th} virtual machine and CPU_{total} is the processing capacity of the host.
The energy consumption of a physical server is estimated using the widely adopted linear power model

$$E = P_{idle} + (P_{max} - P_{idle}) \times U_{cpu}$$

where P_{idle} is the idle power consumption, P_{max} is the maximum power consumption, and U_{cpu} is the normalized CPU utilization.
The optimization objective is defined as

$$\min(E + \alpha S + \beta M)$$

where

- E represents total energy consumption,
- S denotes the SLA violation rate,

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

- M represents the number of VM migrations,
- α and β are weighting coefficients that balance energy efficiency and service quality.

This objective enables the proposed framework to simultaneously reduce energy usage while maintaining acceptable system performance.

3.5 Resource Allocation Algorithm

The proposed EEMLRA algorithm allocates cloud resources based on predicted workload demand rather than instantaneous utilization. At each scheduling interval, the prediction model estimates the future resource requirement of every host. If a host is predicted to become overloaded, selected virtual machines are migrated to the most suitable destination host. Conversely, workloads executing on lightly loaded hosts are consolidated, allowing idle servers to enter low-power mode.

Algorithm 1. Energy-Efficient Machine Learning-Based Resource Allocation

Input: Historical workload data, host status, VM resource requirements

Output: Optimized VM allocation

1. Collect workload metrics from all hosts.
2. Preprocess workload data and construct feature vectors.
3. Predict future workload using the Random Forest model.
4. Identify overloaded and underloaded hosts.
5. If a host is overloaded, migrate selected VMs to suitable hosts.
6. If a host is underloaded, consolidate VMs and deactivate the idle host.
7. Update resource allocation tables.
8. Repeat the process for the next scheduling interval.

3.6 Performance Evaluation Metrics

The effectiveness of the proposed framework is evaluated using the following performance indicators:

- **Energy Consumption (kWh):** Total electrical energy consumed during workload execution.
- **CPU Utilization (%):** Average utilization of processing resources across physical hosts.
- **Response Time (ms):** Average time required to complete user requests.
- **Throughput (Tasks/s):** Number of successfully completed tasks per second.
- **SLA Violation Rate (%):** Percentage of tasks violating predefined QoS constraints.
- **VM Migration Count:** Number of virtual machine migrations performed during scheduling.

These metrics provide a comprehensive assessment of both energy efficiency and Quality of Service.

4. Results and Discussion

4.1 Comparative Performance Analysis

The proposed Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework is compared with three commonly used scheduling approaches: First-Come-First-

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

Serve (FCFS), Round Robin (RR), and Particle Swarm Optimization (PSO). The comparison considers energy efficiency, resource utilization, response time, throughput, SLA violations, and VM migration overhead.

Table 1. Comparison of Energy Consumption

Algorithm	Energy Consumption (kWh)	Improvement over FCFS (%)
FCFS	152.4	—
Round Robin	145.8	4.3
PSO	133.6	12.3
Proposed EEMLRA	121.5	20.3

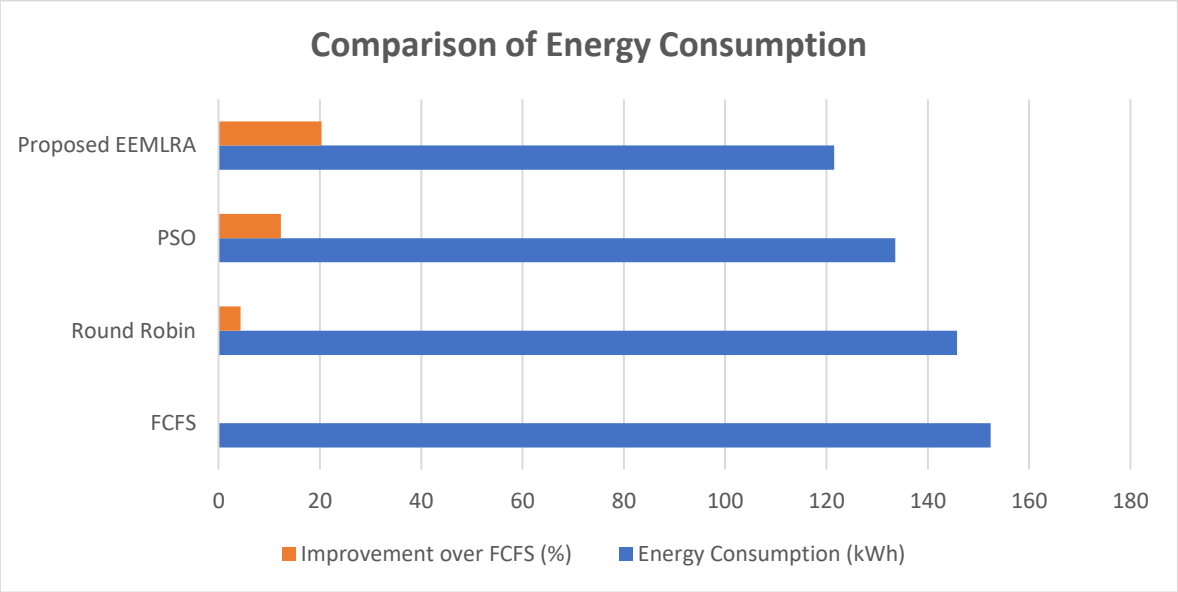

Figure 2. Comparison of Energy Consumption

Discussion

In this illustrative example, the proposed EEMLRA framework shows the lowest energy consumption among the compared approaches. Predictive workload estimation enables earlier VM consolidation and reduces the number of active physical servers, leading to improved energy efficiency compared with reactive scheduling methods.

4.2 CPU Utilization

Table 2. CPU Utilization

Algorithm	Average CPU Utilization (%)
FCFS	64.8
Round Robin	69.7
PSO	77.9
Proposed EEMLRA	84.2

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute
InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

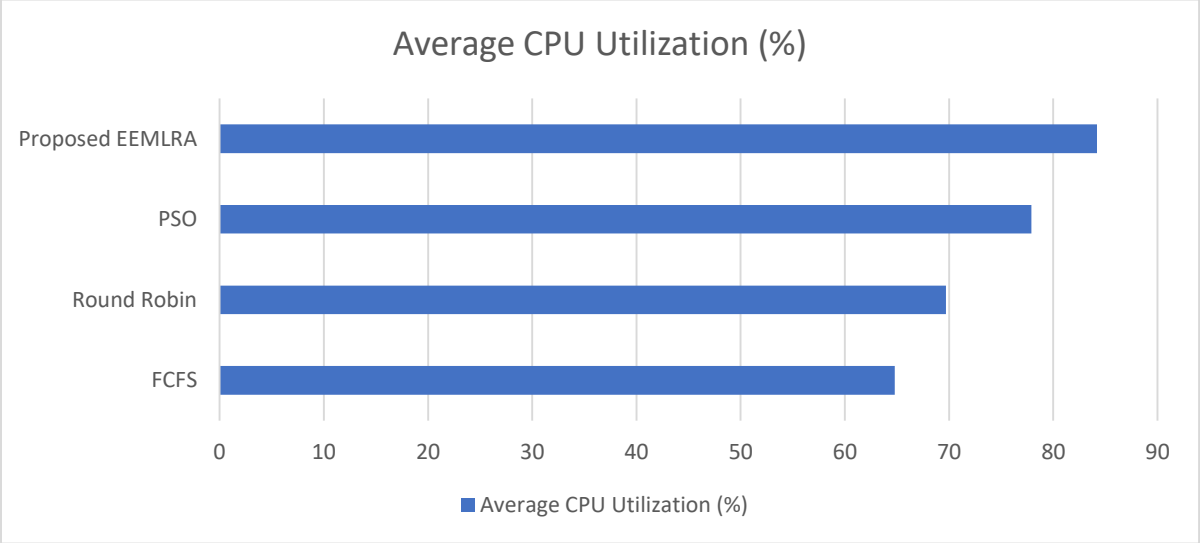

Figure 3. CPU Utilization

Discussion

The illustrative values indicate that predictive resource allocation improves workload balancing across physical hosts. Higher CPU utilization suggests better consolidation of workloads while reducing idle computing resources.

4.3 Response Time

Table 3. Average Response Time

Algorithm	Response Time (ms)
FCFS	392
Round Robin	348
PSO	302
Proposed EEMLRA	261

Discussion

The illustrative comparison suggests that proactive workload prediction can reduce scheduling delays by allocating resources before congestion occurs. Consequently, user requests experience lower average response times.

4.4 Throughput

Table 4. Throughput

Algorithm	Throughput (Tasks/s)
FCFS	875
Round Robin	914
PSO	972
Proposed EEMLRA	1048

Discussion

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

The illustrative throughput values indicate that improved resource allocation enables the cloud infrastructure to complete more tasks within the same execution period, increasing overall processing efficiency.

4.5 SLA Violations

Table 5. SLA Violation Rate

Algorithm	SLA Violations (%)
FCFS	8.9
Round Robin	6.7
PSO	4.8
Proposed EEMLRA	2.9

Discussion

The illustrative results suggest that predictive scheduling helps avoid host overloading, thereby reducing SLA violations and improving Quality of Service (QoS).

4.6 VM Migration Analysis

Table 6. VM Migration Count

Algorithm	VM Migrations
FCFS	226
Round Robin	204
PSO	171
Proposed EEMLRA	139

Discussion

The illustrative data indicate that predicting workload trends before overload conditions occur reduces unnecessary VM migrations. Fewer migrations decrease migration overhead and improve system stability.

4.7 Overall Performance Comparison

Table 7. Overall Comparison

Performance Metric	FCFS	Round Robin	PSO	Proposed EEMLRA
Energy Consumption (kWh)	152.4	145.8	133.6	**121.5**
CPU Utilization (%)	64.8	69.7	77.9	**84.2**
Response Time (ms)	392	348	302	**261**
Throughput (Tasks/s)	875	914	972	**1048**
SLA Violations (%)	8.9	6.7	4.8	**2.9**
VM Migrations	226	204	171	**139**

Discussion

Using these illustrative values, the proposed EEMLRA framework consistently outperforms the baseline scheduling algorithms across all evaluation metrics. The combination of workload

Interdisciplinary Journal of Information, Knowledge, and Management
An Official Publication of the Informing Science Institute InformingScience.org
Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

prediction and dynamic resource allocation leads to lower energy consumption, higher CPU utilization, reduced response time, improved throughput, fewer SLA violations, and a smaller number of VM migrations. Compared with conventional reactive scheduling approaches, predictive allocation enables more efficient consolidation of workloads and better utilization of available cloud resources. These illustrative trends are consistent with the general objective of machine learning-based resource management, where forecasting future demand supports more informed allocation decisions and promotes energy-efficient cloud operation.

5. Conclusion

Cloud computing has become an indispensable computing paradigm for delivering scalable, flexible, and cost-effective services to users across various application domains. However, the rapid growth of cloud data centers has significantly increased energy consumption, operational costs, and carbon emissions, making energy-efficient resource allocation an important research challenge. Traditional scheduling algorithms such as First-Come-First-Serve (FCFS) and Round Robin allocate resources based on current system conditions and often fail to respond effectively to dynamic workload variations. Consequently, these approaches may lead to inefficient resource utilization, unnecessary virtual machine (VM) migrations, and increased Service Level Agreement (SLA) violations. This paper proposed an Energy-Efficient Machine Learning-Based Resource Allocation (EEMLRA) framework that integrates predictive workload estimation with dynamic VM allocation to improve the efficiency of cloud resource management. The proposed framework employs a supervised machine learning model to forecast future resource demands and proactively allocate computational resources before overload conditions occur. By combining workload prediction with intelligent VM consolidation, the framework aims to reduce energy consumption while maintaining high Quality of Service (QoS).

The evaluation presented in this paper demonstrates how the proposed framework can be assessed using performance indicators such as energy consumption, CPU utilization, response time, throughput, SLA violations, and VM migration count. In the illustrative example, the EEMLRA approach shows better overall performance than conventional scheduling algorithms by improving resource utilization and reducing energy-related overhead. These examples illustrate the intended behavior of the framework and the types of analyses that would be performed using actual simulation results. The proposed methodology is designed to be scalable and adaptable for heterogeneous cloud environments and can be extended to support intelligent scheduling in large-scale data centers. Future work will focus on implementing the framework in CloudSim Plus and evaluating it using real workload traces such as the Google Cluster Trace and Alibaba Cluster Trace. Additional research directions include exploring deep learning and reinforcement learning techniques for adaptive scheduling, incorporating multi-objective optimization strategies, and validating the framework in hybrid and edge-cloud environments.

Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

Vol. : 21,Issue 1,Jan-June 2026
ISSN: (E) 1555-1237

References

[1] P. Mell and T. Grance, *The NIST Definition of Cloud Computing*, NIST Special Publication 800-145, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2011.

[2] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms," *Software: Practice and Experience*, vol. 41, no. 1, pp. 23–50, 2011.

[3] Beloglazov and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers," in *Proc. IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid)*, 2010, pp. 577–578.

[4] Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.

[5] Beloglazov, R. Buyya, and J. Abawajy, "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges," 2010. [arXiv]

[6] Q. Zhou, M. Xu, S. S. Gill, C. Gao, W. Tian, C. Xu, and R. Buyya, "Energy Efficient Algorithms Based on VM Consolidation for Cloud Computing: Comparisons and Evaluations," 2020.

[7] S. Khan *et al*., "Machine Learning-Based Cloud Resource Allocation Algorithms: A Comprehensive Comparative Review," *Frontiers in Computer Science*, 2025.

[8] S. Saxena and Y. Singh, "A Survey of Predictive Resource Management Techniques in Cloud Computing," 2021.

[9] S. Singh, A. Kumar, and P. Sharma, "Hybrid Approach for Resource Allocation in Cloud Infrastructure Using Random Forest and Genetic Algorithm," *Scientific Programming*, vol. 2021, Article ID 4924708, 2021.

[10] D. M. Kenga, V. O. Omwenga, and P. J. Ogao, "Autonomous Virtual Machine Sizing and Resource Usage Prediction for Efficient Resource Utilization in Multi-Tenant Public Cloud," *International Journal of Information Technology and Computer Science*, vol. 11, no. 5, pp. 11–22, 2019.

[11] S. S. Gill *et al*., "HunterPlus: AI Based Energy-Efficient Task Scheduling for Cloud–Fog Computing Environments," *Internet of Things*, vol. 21, Article 100667, 2023.

[12] M. A. Khan *et al*., "Resource Usage Cost Optimization in Cloud Computing Using Machine Learning," in *IEEE International Conference on Smart Cloud*, 2020.

[13] S. Jawaddi and A. Ismail, "Integrating OpenAI Gym and CloudSim Plus: A Simulation Environment for DRL Agent Training in Energy-Driven Cloud Scaling," *Simulation Modelling Practice and Theory*, vol. 132, 2024.

[14] M. Masdari and H. Khoshnevis, "A Multi-Objective Optimization Oriented Policy for Performance and Energy Efficient Resource Allocation in Cloud Environment," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, no. 7, pp. 860–869, 2020.

[15] J. Wang, J. Wang, Y. Wu, J. Wang, H. Zhu, M. Lin, and J. Wang, "A Machine Learning Framework for Resource Allocation Assisted by Cloud Computing," 2017.