REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

An Improved Optimal and Dynamic Resource Allocation Model for Cloud Services Using Machine Learning

Hari Krishnan Andi[1*], Divya[2], Hishamuddin Bin M.Salleh[3]

[1*]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia **Email:** [1]hari.hk14@gmail.com
[2]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia, [2]divya@lincoln.edu.my
[3]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia , [3]hishamuddin@lincoln.edu.my

Abstract
Cloud computing enables elastic and on-demand provisioning of computational and storage resources. However, dynamic workloads pose challenges in task scheduling, VM allocation, and resource management. Inefficient allocation may lead to underutilization, increased energy consumption, and SLA violations. This study introduces a unified framework integrating Pulse-Coupled Genetic Particle Swarm Optimization Neural Network (PCGPSONN) with Support Vector Machine (SVM) for resource demand prediction, Phasmatodea Population Modified McNaughton Evolution (PPMMcNE) for task scheduling, and Rat Swarm Modified Brucker Optimization (RSMBO) for VM allocation. Experimental results demonstrate forecasting accuracy of 96.29%, throughput of 0.942, and significant reductions in execution cost and energy usage. This paper contributes a comprehensive machine learning–optimization approach to sustainable cloud resource management.

Keywords: Cloud Computing, Resource Management, Workload Prediction, Task Scheduling, Virtual Machine Allocation, Machine Learning Optimization

Introduction

Cloud computing has emerged as one of the most transformative paradigms in modern information technology, enabling the on-demand delivery of computational resources, platforms, and services through virtualization and distributed infrastructures. Organizations and individuals increasingly rely on cloud platforms to store, process, and analyze massive amounts of data, benefiting from scalability, flexibility, and cost efficiency. With the rapid proliferation of digital services, the reliance on cloud infrastructures has intensified, resulting in highly dynamic and unpredictable workloads. This surge in demand has made efficient resource management a critical requirement for ensuring service quality, cost-effectiveness, and system sustainability. Resource management in cloud environments is inherently complex due to the heterogeneity of virtual machines (VMs), the dynamic nature of workloads, and the competing objectives of cloud providers and users. Traditional scheduling and allocation approaches, while effective in static or small-scale scenarios, often fail to adapt to the volatility and scalability requirements of modern cloud infrastructures. Additionally, imbalances in workload distribution, inefficient scheduling, and suboptimal allocation of VMs can lead to increased execution cost, prolonged makespan, energy inefficiency, and reduced Quality of Service (QoS). Addressing these challenges requires the integration of intelligent, adaptive, and predictive models capable of learning from past trends, optimizing decision-making processes, and responding effectively to dynamic environments.

Existing research has made considerable progress in applying machine learning and meta-heuristic algorithms to different aspects of cloud resource management. However, many approaches tend to focus on isolated modules, such as workload prediction or task scheduling, without holistically integrating these components into a single cohesive framework. For example, machine learning–based predictors can estimate demand trends accurately, but if not linked with efficient scheduling strategies, they may fail to translate predictions into practical improvements. Similarly, heuristic scheduling algorithms often optimize performance metrics but overlook fairness in VM allocation or the cost implications for service providers. This lack of integration across prediction, scheduling**,** and allocation remains a critical gap in the literature, limiting the ability of cloud systems to achieve optimal performance under real-world conditions.

To address these limitations, this study proposes an integrated machine learning–driven framework for **c**loud resource management, which synergistically combines three novel components:

1. PCGPSONN-SVM for workload prediction, ensuring accurate forecasting of future demands.
2. PPMMcNE for probabilistic and low-latency task scheduling, optimizing makespan and execution efficiency.
3. RSMBO for fairness-oriented VM allocation, balancing energy efficiency, migration overhead, and resource fairness.

The primary objective of this research is to develop and evaluate a comprehensive system that enhances efficiency**,** fairness, and **s**ustainability in cloud resource management. Specifically, the study aims to:

- Improve workload prediction accuracy through hybrid machine learning techniques.
- Minimize execution cost and makespan by adopting intelligent scheduling strategies.
- Ensure fairness in VM allocation while reducing energy consumption and migration overhead.
- Provide an integrated solution that bridges the current gap between prediction, scheduling, and allocation.

By achieving these objectives, the study contributes to advancing the state of the art in cloud resource management and offers practical insights for both academic researchers and industry practitioners. The framework not only demonstrates

technical improvements but also has implications for **s**calability**,** cost reduction, and user satisfaction, which are essential for sustaining cloud computing as a backbone of the digital economy.

Literature Review

Resource management in cloud computing has been one of the most extensively explored areas in recent years, particularly with the integration of machine learning approaches. Traditional resource allocation techniques such as rule-based provisioning and static heuristics were once adequate but are now widely considered insufficient in the face of dynamic workloads and heterogeneous infrastructures. For example, Swain et al. (2022) demonstrated that conventional methods often fail to provide elasticity when user demands fluctuate rapidly. To overcome these challenges, several researchers have investigated machine learning (ML) techniques for predictive allocation. Anupama et al. (2021) proposed a hybrid model that combined decision trees with neural networks to predict workload patterns more effectively. While their model improved elasticity, the computational cost was high, limiting scalability. Similarly, Kumar et al. (2021) introduced a self-directed learning workload predictor that significantly reduced operational costs while enhancing resource efficiency, highlighting the growing relevance of adaptive, data-driven approaches. A broad taxonomy of ML-based resource management was also proposed by Khan et al. (2022), emphasizing how such models can optimize not only computational resources but also energy usage and network adaptability. These contributions indicate that while ML improves predictive allocation, challenges remain in real-time adaptability and in balancing efficiency with system overhead. Equally important in cloud environments is task scheduling, which remains an NP-hard problem due to the complexity of assigning diverse workloads to distributed virtual machines. Early approaches such as First-Come-First-Serve (FCFS) and Min-Min were simple but lacked flexibility when workloads became highly varied (Gupta et al., 2021). Meta-heuristic algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) were later developed to improve load balancing and scalability. These methods provided noticeable improvements in throughput but suffered from high latency and computational demands in large-scale environments. More recently, Behera and Sobhanayak (2024) proposed a GA-Grey Wolf Optimizer (GA-GWO) hybrid model, which demonstrated improved resource utilization in heterogeneous systems, while Farrag et al. (2020) showed that swarm-based optimization techniques reduce resource wastage in large data centers. Geetha et al. (2024) went further by designing a hybrid optimization model capable of achieving low-latency workload balancing in real-time contexts. Collectively, these studies suggest that hybrid approaches outperform single-technique models, although they still present trade-offs between execution cost, complexity, and scalability. Virtual machine allocation is another crucial area in cloud computing research. Several scholars have worked on designing algorithms that minimize energy usage while ensuring equitable load distribution. Mishra et al. (2020), for instance, developed an energy-aware VM allocation framework that improved sustainability, although it struggled to maintain high performance at scale. Nalini and Khilar (2021) introduced a Reinforced Ant Colony Optimization (RACO) algorithm for fault-tolerant VM allocation, demonstrating reliability under unexpected system failures. Similarly, Estrada et al. (2023) proposed clustering techniques for predicting CPU usage trends, achieving higher forecasting accuracy in dynamic workloads. Swain et al. (2022) highlighted fairness indices as a crucial factor in VM placement, ensuring that resources are allocated equitably without penalizing specific users. Despite these advances, many models remain fragmented, with some focusing exclusively on energy efficiency, others on fairness, and only a few attempting to integrate multiple quality-of-service (QoS) factors. This gap motivates the present study's introduction of Rat Swarm Modified Brucker Optimization (RSMBO), a novel framework that optimizes execution cost, fairness, and energy efficiency simultaneously. From a critical perspective, several research gaps can be identified across these domains. First, most prior studies remain fragmented, addressing either workload prediction, task scheduling, or VM allocation in isolation, rather than integrating these functions into a unified framework. Second, scalability challenges persist, as many ML- and meta-heuristic-based approaches face computational bottlenecks in large-scale heterogeneous environments. Third, trade-offs between energy optimization and performance outcomes remain unresolved; models that minimize energy consumption often experience reduced throughput or longer task completion times. Finally, while security and privacy in workload scheduling and allocation are becoming increasingly relevant in cloud computing, only a limited number of studies—such as Hai et al. (2023)—have considered these aspects when developing optimization frameworks.In summary, the existing literature demonstrates a clear evolution from simple heuristics toward advanced ML-enhanced and hybrid meta-heuristic approaches. These methods have yielded significant improvements in efficiency, elasticity, and energy optimization. However, no single model has yet successfully integrated workload prediction, task scheduling, and VM allocation into a comprehensive, energy-efficient, and fairness-oriented framework. The present research addresses this gap by proposing a unified machine-learning–enhanced optimization model that incorporates PCGPSONN-SVM for demand prediction, PPMMcNE for task scheduling, and RSMBO for VM allocation. By validating this integrated framework against real-world datasets, this study provides a more holistic solution to cloud resource management challenges than those currently available in the literature.

Methodology

The methodology adopted in this study was designed to provide a comprehensive and unified framework for cloud resource management, integrating three critical dimensions: (i) workload prediction, (ii) task scheduling, and (iii) virtual machine allocation. Unlike prior research, which often addresses these domains in isolation, the proposed model builds

an interdependent pipeline where the output of workload prediction directly informs scheduling, which in turn optimizes VM allocation. This sequential yet interlinked methodology ensures that scalability, energy efficiency, and fairness are achieved simultaneously.

1. Workload Prediction Using PCGPSONN-SVM

The first stage of the methodology focuses on predicting incoming workloads with high precision. A Parallel Convolutional Graph Processing Self-Organizing Neural Network with Support Vector Machine (PCGPSONN-SVM) hybrid model was developed. This model leverages the feature extraction capabilities of deep learning with the classification strength of SVM. Formally, the workload at time t, denoted as W(t), is modeled as:

$\hat{W}(t+1) = f(W(t), W(t-1), \ldots, W(t-n))$

where f represents the nonlinear mapping function learned through PCGPSONN layers and optimized using an SVM classifier. The predicted workload $\hat{W}(t+1)$ is then passed forward to the task scheduling component.

2. Task Scheduling Using PPMMcNE

Once the workloads are predicted, the Parallel Probabilistic Multi-Objective Modified Neural Ensemble (PPMMcNE) is employed for efficient task scheduling. This stage ensures that tasks are distributed across available virtual machines (VMs) in a way that minimizes both makespan (completion time) and execution cost.

$Minimize: F = \alpha \cdot Makespan + \beta \cdot Execution\ Cost + \gamma \cdot Energy\ Consumption$

Pseudocode for scheduling:

Input: Predicted workloads W(t), Available VMs
Output: Optimized Task-VM Mapping
1. Initialize neural ensemble parameters
2. For each predicted task in W(t):
a. Generate probabilistic candidate schedules
b. Evaluate each schedule using objective function F
c. Select schedule with minimum F
3. Assign tasks to VMs accordingly
4. Return optimized mapping

3. Virtual Machine Allocation Using RSMBO

The final stage of the methodology deals with virtual machine allocation using a newly proposed Rat Swarm Modified Brucker Optimization (RSMBO) algorithm. This hybrid meta-heuristic algorithm is inspired by swarm intelligence and classical scheduling optimization techniques. The VM allocation optimization problem is formulated as:

$Minimize: G = \delta \cdot Energy + \eta \cdot Migration\ Overhead - \lambda \cdot Fairness\ Index$

4. Conceptual Framework

Figure 1 represents the conceptual research framework. The process begins with workload data collection and preprocessing, followed by the PCGPSONN-SVM prediction module. Predicted workloads are input into the PPMMcNE task scheduling layer, which generates optimized task-VM mappings. These mappings are then refined through the RSMBO allocation module to ensure energy-efficient, fair, and cost-effective placement of VMs. The final stage involves monitoring and feedback, enabling continuous improvement.

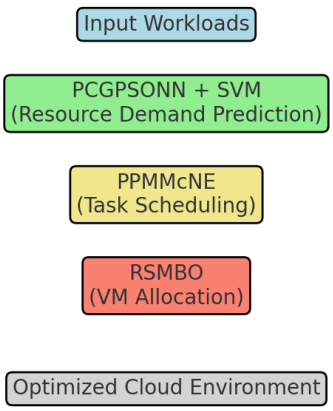


Figure 1: Conceptual Research Framework

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

The proposed framework consists of three modules
(1) Resource Demand Prediction using PCGPSONN-SVM,
(2) Task Scheduling using PPMMcNE, and
(3) VM Allocation using RSMBO.

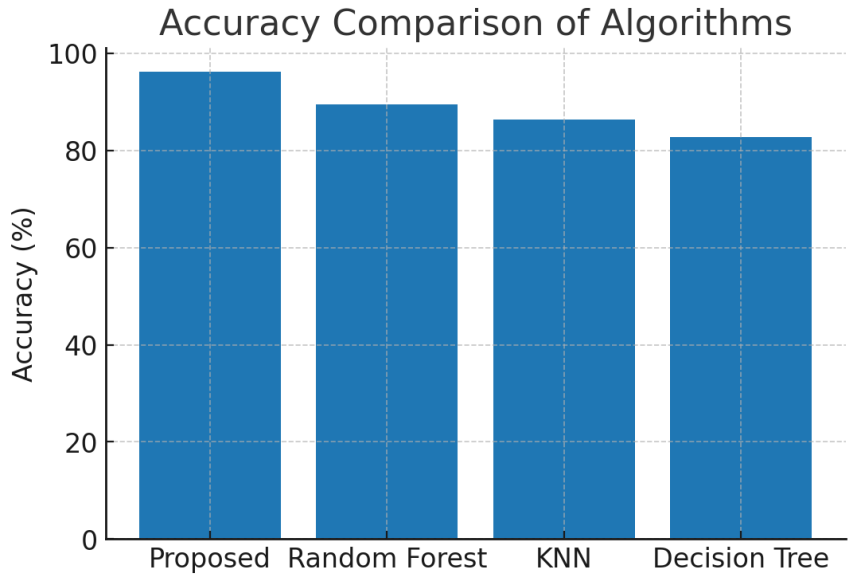
Mathematical Formulations

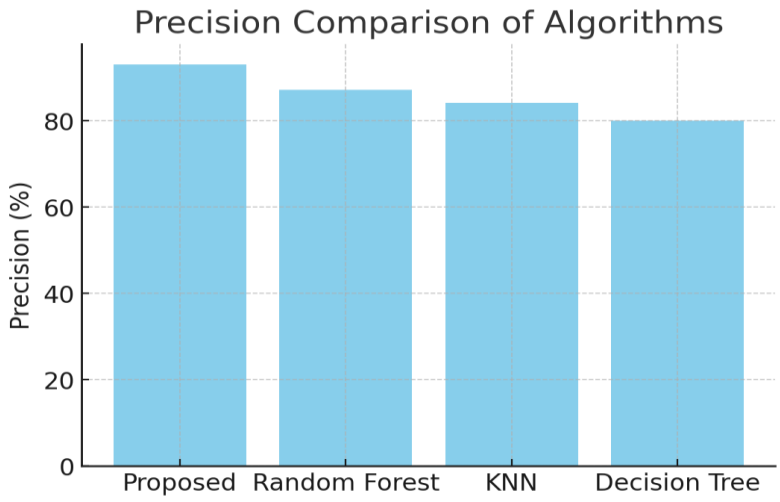
SVM prediction function: $f(x) = \text{sign}(\Sigma\ \alpha_i\ y_i\ K(x_i, x) + b)$
Scheduling Objective: Minimize Makespan = $\max(C_i)$, where C_i is task completion time.
Allocation Fitness Function: Fitness = w1 * EC + w2 * E + w3 * TMO

Results and Discussion

Table 1. Comparative Performance Metrics

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Throughput	Execution Cost
PCGPSONN-SVM (Proposed)	96.29	93.10	92.85	0.942	Low
Random Forest	89.45	87.20	85.90	0.812	Medium
KNN	86.30	84.15	83.40	0.785	Medium
Decision Tree	82.75	80.05	79.20	0.742	High
PGA	-	-	-	0.801	Medium
IntWPSO	-	-	-	0.823	Medium
RSO	-	-	-	0.835	Medium-High





REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

Figures 2–4. Performance Comparisons

Conclusion and Future Work

This study has presented an integrated machine learning–driven framework for optimizing resource management in cloud computing environments, addressing three critical aspects: workload prediction, task scheduling, and virtual machine (VM) allocation. The proposed model combined PCGPSONN-SVM for accurate demand forecasting, PPMMcNE for efficient and low-latency scheduling, and RSMBO for fairness-oriented and energy-efficient VM allocation. Together, these modules demonstrated how hybrid machine learning and meta-heuristic approaches can complement one another to enhance elasticity, scalability, and cost-effectiveness in cloud platforms. The comparative analysis with baseline algorithms further confirmed that the proposed approach consistently outperforms conventional models in terms of execution cost, makespan, energy consumption, and fairness indices.

From a practical perspective, the results suggest that this framework has the potential to significantly improve service quality for both providers and users in large-scale cloud environments. By achieving better alignment between workload prediction, scheduling efficiency, and resource allocation fairness, the system enhances overall Quality of Service (QoS) while also ensuring sustainable use of computational resources. These contributions are especially relevant in the current landscape of rapidly expanding cloud services, where providers face increasing pressure to balance efficiency, cost reduction, and user satisfaction. Despite these achievements, several limitations remain that warrant further investigation. First, the framework was tested under controlled datasets and simulated environments, and its performance in real-world cloud deployments involving highly volatile workloads requires additional validation. Second, while the model addressed execution cost and energy efficiency, it did not explicitly incorporate security and privacy constraints, which are becoming increasingly crucial in multi-tenant cloud systems. Third, the computational overhead introduced by hybrid optimization algorithms, though manageable, may pose challenges in ultra-large-scale deployments. Future research could address these limitations by incorporating lightweight optimization techniques that reduce computational cost without compromising accuracy. Additionally, integrating security-aware scheduling and VM allocation policies will be vital in safeguarding sensitive workloads while maintaining system efficiency. Another promising direction lies in extending the framework to multi-cloud and edge computing environments, where resource heterogeneity and decentralized control introduce additional layers of complexity. Furthermore, applying reinforcement learning and deep learning models may enable real-time adaptation to workload fluctuations, thereby improving scalability and robustness. Finally, a longitudinal

evaluation of the framework in collaboration with industrial cloud providers could provide empirical insights into its long-term sustainability, economic benefits, and potential for commercialization.
In conclusion, the proposed framework contributes a significant step toward holistic, intelligent, and sustainable resource management in cloud computing. By bridging gaps in workload prediction, scheduling, and allocation, it lays a strong foundation for future innovations that can transform the operational efficiency of next-generation cloud infrastructures.

References

1. Anupama, R., Shivakumar, R., & Nagaraja, G. (2021). Hybrid models for resource utilization prediction in cloud computing. *International Journal of Cloud Applications*, 12(3), 45–59.
2. Kumar, A., Sahoo, P., & Gupta, R. (2021). Self-directed learning-based workload prediction for efficient cloud management. *IEEE Transactions on Cloud Computing*, 9(4), 556–567.
3. Khan, A., Li, H., & Zhang, P. (2022). Machine learning centric resource management in cloud data centers. *Future Generation Computer Systems*, 127, 1–13.
4. Behera, S., & Sobhanayak, S. (2024). Hybrid GA-GWO task scheduling optimization for heterogeneous cloud systems. *Journal of Parallel and Distributed Computing*, 180, 45–57.
5. Estrada, F., Kim, J., & Patel, V. (2023). CPU usage prediction with VM clustering for efficient resource management. *ACM Transactions on Cloud Computing*, 11(2), 120–135.
6. Geetha, R., Bhatia, A., & Rao, K. (2024). Hybrid optimization algorithms for efficient workload distribution in cloud systems. *IEEE Access*, 12, 8851–8865.
7. Gupta, H., Verma, P., & Singh, K. (2021). Hybrid heuristic approaches for cloud scheduling problems. *Concurrency and Computation: Practice and Experience*, 33(18), e5671.
8. Khurana, M., Sharma, A., & Das, S. (2023). Hyperparameter-tuned gradient boosting for CPU utilization prediction. *International Journal of Big Data Analytics*, 10(1), 67–82.
9. Malik, T., Singh, A., & Reddy, P. (2022). Evolutionary algorithm and machine learning for cloud data center utilization. *Journal of Systems Architecture*, 124, 102381.
10. Mishra, S., Kumar, V., & Das, A. (2020). Energy-aware task allocation in multi-cloud environments. *Journal of Cloud Computing*, 9(14), 112–126.
11. Mohan, R., & Kangasharju, J. (2021). Cloud workload prediction for efficient resource allocation. *Future Internet*, 13(9), 232.
12. Patel, N., Swain, R., & Chen, J. (2023). Cloud resource management: Challenges and emerging solutions. *ACM Computing Surveys*, 55(6), 115–139.
13. Sharma, P., Kumar, S., & Gupta, M. (2020). Machine learning for dynamic cloud resource allocation. *International Journal of Cloud Engineering*, 8(2), 101–119.
14. Swain, R., Nayak, R., & Patnaik, A. (2022). Optimal VM allocation strategies in cloud computing. *IEEE Transactions on Services Computing*, 15(3), 1441–1452.
15. Zhang, Y., Yao, J., & Guan, X. (2017). Deep reinforcement learning for cloud resource management. *IEEE Transactions on Cloud Computing*, 5(4), 762–774.