

A Comparative Review of AI-Driven Predictive Resource Management for Cost, Performance, and Energy Optimization in Cloud Computing

Priyanka Sunil Aher, Jyoti Anil Mahatme

¹Department of Computer Science, [K. R. T. Art's, B. H. Commerce and A. M. Science (KTHM) College, Nashik], Maharashtra, India

²Department of Computer Science, [K. R. T. Art's, B. H. Commerce and A. M. Science (KTHM) College, Nashik], Maharashtra, India

Email: [priyankaaher@kthmcollege.ac.in, jyotimahatme@kthmcollege.ac.in]

Abstract

Cloud computing resource management must continuously respond to fluctuating workloads while balancing performance, cost, scalability, service-level agreement (SLA) requirements, and energy consumption. Reactive threshold-based allocation and autoscaling can lead to over-provisioning, under-provisioning, delayed adaptation, and inefficient resource utilization. This paper presents a secondary-data comparative review of artificial intelligence (AI) and machine learning (ML) approaches for predictive workload management, dynamic resource allocation, autoscaling, cost optimization, and energy-efficient cloud operation. Fifteen supplied references were examined; the duplicate entry was consolidated, and reported methods, datasets, objectives, evaluation measures, and quantitative outcomes were compared. The reviewed evidence shows that workload forecasting, deep learning, reinforcement learning, and hybrid AI approaches can improve resource-management decisions, while published results remain difficult to compare because studies use different datasets, simulators, baselines, and metrics. Reported examples include up to 25% operational-cost reduction in a multi-cloud framework, substantial workload-prediction error reductions using neural networks with adaptive differential evolution, and large improvements in makespan, cost, and energy metrics in selected ML resource-allocation studies. The review identifies a major gap: prediction, allocation, autoscaling, cost, and energy objectives are often studied separately or evaluated under inconsistent conditions. A unified, benchmark-driven and multi-objective predictive resource-management framework is therefore proposed as a future research direction.

Keywords: Cloud Computing, Artificial Intelligence, Machine Learning, Workload Prediction, Resource Allocation, Autoscaling.

1. INTRODUCTION

Cloud computing provides elastic access to computing, storage, and networking resources, but the elasticity that makes cloud systems attractive also creates a difficult resource-management problem. Workloads can change rapidly because of user demand, application behavior, seasonal patterns, and bursty service requests. If resources are provisioned above demand, providers incur unnecessary cost and energy consumption. If resources are provisioned below demand, applications may experience latency, lower throughput, and SLA violations. Recent literature therefore increasingly shifts from reactive management toward predictive and adaptive resource management.

The supplied literature covers this transition from several perspectives. Kumar and Singh proposed neural-network workload prediction optimized with self-adaptive differential evolution and evaluated it using NASA and Saskatchewan HTTP traces [3]. Adeyinka combined workload forecasting with long short-term memory (LSTM), reinforcement learning, and decision-tree techniques for dynamic multi-cloud allocation and reported operational-cost reductions of up to 25% [1]. Recent surveys and comparative studies examine machine-learning-based autoscaling, resource allocation, predictive management, cost optimization, and energy-aware management [2], [4], [6]–[8].

The purpose of this paper is not to claim new experimental performance from the supplied studies. Instead, it performs a structured secondary-data comparison of the 15 supplied reference entries, consolidates the duplicate title, identifies common methods and evaluation measures, and derives research gaps that can guide a future experimental framework. This distinction is important because reported percentages from different papers cannot be treated as a single statistically comparable dataset without common experimental conditions.

2. OBJECTIVES OF THE STUDY

- To review AI and ML techniques used for predictive cloud resource management.
- To compare workload prediction, resource allocation, autoscaling, cost optimization, and energy-management approaches reported in the supplied literature.
- To identify common datasets, evaluation metrics, platforms, and reported performance indicators.
- To determine research gaps that limit direct comparison and practical deployment.
- To propose a research direction for an integrated predictive resource-management framework.

3. RESEARCH METHODOLOGY

This study uses secondary research data taken from the below 15 references. The review process consists of five stages: (1) identification of the supplied publications; (2) classification by problem area and AI technique; (3) extraction of datasets, algorithms, objectives, and reported metrics; (4) qualitative and quantitative comparison of reported outcomes; and (5) synthesis of research gaps and future directions.

The comparison is descriptive rather than a meta-analysis. Results reported by different authors are retained in their original context because the studies use different workloads, experimental platforms, baselines, parameter settings, and definitions of improvement. Consequently, a value such as 25% cost reduction is reported as a finding of its source study and is not presented as a universal performance guarantee.

4. LITERATURE REVIEW

4.1 Workload prediction. Kumar and Singh [3] addressed workload forecasting using an artificial neural network combined with self-adaptive differential evolution. The adaptive evolutionary component learned suitable mutation and crossover behavior, and experiments used NASA and Saskatchewan HTTP traces. The study reported minimum prediction errors of 0.013 and 0.001 for the two datasets and emphasized that accurate prediction can support better scaling decisions.

4.2 Predictive allocation and multi-cloud management. Adeyinka [1] integrated workload forecasting with LSTM, reinforcement learning, and decision-tree-based techniques for multi-cloud resource allocation. The framework continuously monitors workload patterns and dynamically

changes provisioning. The paper reports improved CPU and memory utilization, increased SLA compliance, and up to 25% lower operational cost.

4.3 Autoscaling. Machiraju, Kumar, and Sharma [2] surveyed 60 ML-based autoscaling studies published between 2015 and 2025 and organized them using dimensions including goal, decision logic, scaling mode, control scope, and deployment. The review shows that autoscaling is a fragmented problem involving different objectives and platforms, and highlights challenges such as telemetry/actuation delay, hybrid scaling complexity, and joint optimization of cost, service objectives, and energy.

4.4 Comparative resource allocation. Bodra and Khairnar [4] compared 10 ML algorithms across deep reinforcement learning, neural networks, traditional ML-enhanced methods, and multi-agent systems. Their synthesis reports that hybrid AI architectures often outperform single-method approaches. One cited result, ATSLA3C, achieved 70.49% makespan reduction, 77.42% cost optimization, and 74.24% energy-consumption improvement in its evaluated setting.

4.5 Integrated ML resource management. Haroon [5] proposed a framework in which Random Forest performs workload prediction and a Deep Q-Learning agent handles resource management, scheduling, and autoscaling. This illustrates a useful architectural separation between prediction and decision-making.

4.6 Predictive management reviews. Yadav and Yadav [6] review supervised, unsupervised, semi-supervised, deep-learning, and reinforcement-learning approaches for workload prediction, VM allocation, autoscaling, load balancing, energy optimization, and SLA mitigation. Thafzy [7] similarly surveys ML integration across container placement, job scheduling, and multi-resource scheduling. Ramamoorthi [8] emphasizes dynamic allocation, predictive analytics, adaptive scaling, energy optimization, and multi-tenant management.

4.7 Cost and optimization. Choudhary [9] discusses an implementation framework for AI-powered predictive analytics covering capacity provisioning, cost management, performance enhancement, and energy efficiency. Polu [12] focuses on predictive cost optimization, anomaly detection, automated right-sizing, FinOps, and multi-cloud workload placement. Olaoye [13] examines AI's impact on cloud cost optimization and resource management. Gadekar and Dhokare [14] discuss AI-based predictive analysis, scaling, load balancing, and resource optimization.

4.8 Multi-cluster and energy efficiency. Punniyamoorthy et al. [11] describe an AI-driven multi-cluster framework using predictive learning, policy-aware decision-making, and continuous feedback to coordinate resource management across clusters. Subhashree et al. [15] propose AI-based dynamic resource management for energy-efficient cloud data centres, combining workload prediction, VM placement, and dynamic resource allocation while maintaining QoS.

5. COMPARATIVE ANALYSIS

Ref.	Primary focus	AI/ML techniques	Data / evaluation	Reported outcome
[1]	Multi-cloud allocation	LSTM, RL, decision trees	Workload forecasting; multi-cloud framework	Up to 25% operational-cost reduction; improved utilization/SLA
[2]	ML autoscaling	Supervised ML, DL, RL	60-study survey	Identifies fragmentation and multi-objective gaps
[3]	Workload prediction	ANN + self-adaptive DE	NASA & Saskatchewan HTTP traces	Minimum prediction error 0.013 and 0.001
[4]	Resource allocation comparison	10 ML/AI algorithms	Comparative review	Selected studies report >70% improvements in

				some metrics
[5]	Dynamic balancing	Random Forest + DQN	CloudSim-oriented framework	Targets cost, utilization, SLA and autoscaling
[6]	Predictive management	ML/DL/RL families	Review of datasets/metrics/platforms	Identifies heterogeneity, interpretability and real-time gaps
[7]	ML resource management	Supervised/unsupervised/RL	Literature review	Highlights allocation, scheduling and multi-resource management
[8]	AI resource management	ML, RL, optimization	Review	Focuses allocation, energy and multi-tenancy
[9]	Predictive optimization	AI/ML predictive analytics	Technical framework/case evidence	Covers cost, performance, capacity and energy
[11]	Multi-cluster optimization	Predictive learning + feedback	Prototype/framework	Improved resource efficiency and stabilization

				ation
[1 2]	Cloud cost optim ization	Predictive analytics, ML, RL, FinOps	Conceptual/i ndustry- oriented	Targets right- sizing, anomal y detecti on and cost control
[1 3]	AI and cloud cost mana geme nt	AI/ML approaches	Conceptual/r eview analysis	Links AI to cost and resourc e- manag ement decisio ns
[1 4]	Cloud optim ization	ML, predictive analytics	Review/illus trative analysis	Focus s efficien cy, cost and scaling
[1 5]	Energ y- effici ent data centre s	ML, workload prediction, VM placement	Simulated cloud environment	Report s energy savings while maintai ning QoS

6. KEY FINDINGS FROM THE COMPARISON

- Workload prediction is the foundation of proactive resource management. ANN, LSTM, ensemble learning and other predictive models are repeatedly used to estimate future demand.
- Reinforcement learning is mainly valuable after prediction, where an agent can choose allocation, scheduling, or scaling actions under changing conditions.
- Hybrid architectures are promising because forecasting and decision-making are different tasks; a single model does not necessarily need to perform both.
- Cost, performance, SLA compliance, resource utilization, and energy are conflicting objectives. Optimizing one metric alone can produce undesirable trade-offs.

- Multi-cloud and multi-cluster environments add heterogeneity, placement, interoperability, and coordination challenges that are not fully addressed by single-cloud studies.
- Published results are difficult to compare directly because datasets, simulators, baselines, workloads, metrics, and experimental scales differ.

7. RESEARCH GAPS

Gap	Evidence from literature	Need for future research
Separated prediction and control	Many studies focus primarily on prediction or allocation/autoscaling.	Integrate forecasting, allocation and scaling in one closed loop.
Inconsistent evaluation	Reviews identify different datasets, platforms, baselines and metrics.	Use common benchmarks and standardized experimental protocols.
Limited real-world validation	Recent systematic evidence reports heavy reliance on simulation.	Validate models using real cloud traces and, where possible, real orchestration platforms.
Multi-objective optimization	Cost, SLA, performance and energy are often evaluated separately.	Use a unified multi-objective formulation.
Sustainability gap	Energy/carbon awareness is less common than performance/cost optimization.	Include energy and carbon intensity as decision variables.
Model overhead and explainability	Complex DL/RL models may add training/inference cost and reduce interpretability.	Measure model overhead and use explainable/adaptive models where feasible.
Dynamic model drift	Cloud workloads change over time.	Use online learning, drift detection or periodic retraining.
Cross-cloud interoperability	Multi-cloud and multi-cluster systems have	Develop provider-neutral resource

	heterogeneous resources and policies.	abstractions and coordinated policies.
--	---------------------------------------	--

8. PROPOSED RESEARCH FRAMEWORK

Based on the identified gaps, the following framework is proposed as a future experimental extension rather than as a result claimed by this review. Historical and real-time workload traces are first cleaned and transformed into time-series features. A forecasting module predicts CPU, memory, request rate, or task demand for a future horizon. A decision module then converts the forecast into resource-allocation and autoscaling actions. A monitoring layer measures cost, utilization, response time, SLA violations, energy consumption, and prediction error. These observations are fed back for model updating.

The conceptual control loop is: **Workload Data** → **preprocessing** → **Workload Forecasting** → **Resource Demand Estimation** → **Intelligent Allocation** → **Autoscaling** → **Cloud Execution** → **Monitoring** → **Feedback**. A practical implementation can compare Random Forest, ANN and LSTM for forecasting, followed by threshold-based allocation, reinforcement learning, or a multi-objective optimizer. CloudSim Plus can be used for controlled experiments, while real cloud traces can improve external validity.

A multi-objective decision function can be represented as: $F = w_1C + w_2T + w_3S + w_4E - w_5U$, where C is cost, T is response/execution time, S is SLA-violation penalty, E is energy consumption, U is resource utilization, and w_i are configurable weights. The optimization objective is to minimize F while satisfying capacity and QoS constraints. This formulation should be calibrated experimentally rather than assigning arbitrary weights.

9. PERFORMANCE METRICS

Prediction metrics should include mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R^2 where appropriate. Resource-management metrics should include CPU and memory utilization, response time, execution time, makespan, throughput, number of scaling actions, SLA violations, operational cost, energy consumption, and—where data are available—carbon intensity. The same dataset split, workload horizon, baseline

policies, and measurement definitions should be used for all compared models.

10. DISCUSSION

The literature indicates a clear movement from static and reactive resource management toward predictive and adaptive cloud management. The strongest architectural pattern is not necessarily one particular algorithm but the combination of forecasting, decision-making, and continuous feedback. Deep learning is attractive for complex temporal patterns, while Random Forest and other conventional models can be useful when data are structured and computational overhead must remain low. Reinforcement learning is particularly suitable when resource decisions form a sequential control problem.

However, numerical improvements reported in different studies should not be merged into a single ranking. For example, a 25% cost reduction, a prediction error of 0.001, and a 70% makespan reduction measure different outcomes under different experimental conditions. A credible future experiment should therefore reproduce several algorithms on the same workload traces and use identical baselines and metrics.

A 2026 systematic review beyond the supplied list similarly reports substantial but highly variable AI improvements, while finding that 70% of evaluations were simulation-based and only a small fraction incorporated carbon-aware scheduling. This independently supports the gaps identified from the supplied references and strengthens the case for standardized, sustainable, and production-oriented evaluation.

11. LIMITATIONS

This study is limited by the supplied reference set and by the secondary nature of the extracted data. The papers use different experimental settings, so the comparison cannot establish statistical superiority of one algorithm over another. Some supplied titles also have incomplete bibliographic information in the material available for verification. Reference 9 and reference 10 are duplicates and were consolidated. Before journal submission, all references should be checked against the original PDFs and DOI metadata.

12. CONCLUSION

This paper reviewed and compared AI-driven approaches for predictive workload management, dynamic resource allocation, autoscaling, cost optimization, multi-cloud coordination, and energy-efficient cloud operation. The literature consistently indicates that AI and ML can move resource management from reactive provisioning toward proactive decision-making. Workload prediction provides the information needed for anticipation, while reinforcement learning, optimization, and policy-based mechanisms can convert predictions into allocation and scaling actions. The main research gap is the absence of a consistently evaluated, integrated framework that simultaneously considers prediction accuracy, cost, performance, SLA compliance, resource utilization, and energy or carbon impact. Future work should therefore focus on benchmark-driven experiments using common datasets, multiple ML baselines, standardized metrics, multi-objective optimization, online adaptation, and real-world cloud validation. Such a framework can provide stronger evidence for practical deployment of intelligent and sustainable cloud resource management.

REFERENCES

- [1] A. Adeyinka, "Dynamic resource allocation using AI-driven workload forecasting in multi-cloud environments," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 3188–3198, 2024, doi: 10.30574/wjarr.2024.23.1.2178.
- [2] V. S. Machiraju, V. Kumar, and S. Sharma, "ML-Based Autoscaling for Elastic Cloud Applications: Taxonomy, Frameworks, and Evaluation," *Mathematical and Computational Applications*, vol. 31, no. 2, art. 49, 2026, doi: 10.3390/mca31020049.
- [3] J. Kumar and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," *Future Generation Computer Systems*, vol. 81, pp. 41–52, 2018, doi: 10.1016/j.future.2017.10.047.
- [4] D. Bodra and S. Khairnar, "Machine learning-based cloud resource allocation algorithms: a comprehensive comparative review," *Frontiers in Computer Science*, vol. 7, art. 1678976, 2025, doi: 10.3389/fcomp.2025.1678976.
- [5] F. Haroon, "Intelligent Resource Allocation in Cloud Computing Environments: Leveraging Machine Learning for Dynamic Workload Balancing, Cost Efficiency, and Performance Optimization," *Annual Methodological Archive Research Review*, vol. 3, no. 4, pp. 526–549, 2025, doi: 10.63075/gdxsyd86.
- [6] R. Yadav and J. S. Yadav, "A Comprehensive Review of Machine Learning Approaches for Predictive Resource Management in Cloud Computing Environment," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 6, pp. 347–354, 2025, doi: 10.32628/CSEIT2511656.
- [7] Thafzy V. M., "A Review on the Integration of Machine Learning in Cloud Computing Resource Management," *International Journal of Scientific Research and Technology*, vol. 2, no. 1, pp. 18–21, 2025, doi: 10.5281/zenodo.14592545.
- [8] V. Ramamoorthi, "AI-Driven Resource Management in Cloud Computing: A Review," *International Journal of Scientific Research in Science and Technology*, vol. 11, no. 2, pp. 683–696, 2024, doi: 10.32628/IJSRST24113260.
- [9] S. K. Choudhary, "AI-Powered Predictive Analytics for Dynamic Cloud Resource Optimization: A Technical Implementation Framework," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 1267–1275, 2025, doi: 10.32628/CSEIT251112122.
- [10] V. Punniyamoorthy, A. K. Agarwal, B. Kumar, A. Mazumder, K. Kannan, and S. Saha, "AI-Driven Cloud Resource Optimization for Multi-Cluster Environments," *arXiv*, 2025, arXiv:2512.24914.
- [11] O. R. Polu, "AI-Enhanced Cloud Cost Optimization Using Predictive Analytics," *International Journal of Artificial Intelligence Research and Development*, vol. 3, no. 1, pp. 74–88, 2025, doi: 10.34218/IJAIRD_03_01_006.
- [12] G. Olaoye, "The Impact of AI on Cloud Cost Optimization and Resource Management," *SSRN*, posted Apr. 7, 2025.
- [13] S. H. Gadekar and K. K. Dhokare, "AI-Driven Cloud Optimization: Enhancing Efficiency and Cost-Effectiveness," *International Journal of Novel Research and Development*, vol. 9, no. 12, Dec. 2024.
- [14] S. D. C., A. Patil, S. Avareddy, J. R. M., and A. Baig, "Energy-Efficient Cloud Data Centres Using AI-Based Dynamic Resource Management," *Lex Localis – Journal of Local Self-Government*, vol. 23, no. S6, pp. 568–574, 2025, doi: 10.52152/801811.