iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81468

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

A Survey on AI-Based Cloud Cost Management Strategies for SaaS

Anant Kumar Sah[1], Dr. Monika Bhatnagar[2]
[1]*Computer Science and Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal*
[2]*Computer Science and Engineering, Oriental Institute of Science and Technology, Bhopal*

Abstract: ***Cloud cost management is a significant challenge for Software as a Service (SaaS) applications due to dynamic workloads, multi-tenancy, and complex cloud provider pricing models. Traditional approaches such as static provisioning or manual scaling are often insufficient. Artificial Intelligence (AI) techniques—including predictive analytics, reinforcement learning, and optimization algorithms—offer adaptive and intelligent solutions for reducing cloud expenditures while maintaining service-level agreements (SLAs). This paper surveys recent research on AI-driven cloud cost management strategies for SaaS applications, highlighting techniques, applications, challenges, and future directions.***
Keywords: SaaS, Cloud Computing, Cost Optimization, Artificial Intelligence, Reinforcement Learning, Predictive Analytics.

I. INTRODUCTION

Software as a Service (SaaS) has revolutionized the way organizations access and utilize software applications by providing them over cloud infrastructures without the need to manage on-premise resources. SaaS enables businesses to scale their applications rapidly, reduces upfront capital expenditure on hardware and software, and supports remote access and collaboration [1], [4], [6]. With the growing adoption of SaaS platforms across industries, organizations increasingly rely on cloud environments to handle highly dynamic workloads, which can fluctuate unpredictably depending on user demand, seasonal trends, or application-specific events.

However, while SaaS reduces initial infrastructure costs, the operational expenses associated with cloud services continue to rise due to several factors. Multi-tenant architectures, where resources are shared among different clients, complicate the allocation of computational and storage resources, often leading to over-provisioning or underutilization. Furthermore, the pricing models of cloud service providers—such as pay-as-you-go, reserved instances, spot instances, and data transfer fees—add another layer of complexity to cost management. These challenges make it difficult for SaaS providers to predict and control expenditures while simultaneously ensuring high-quality service delivery [1], [4], [6].

Traditional cost management techniques, such as static resource allocation, rule-based autoscaling, and the use of reserved instances, provide a foundation for cost control but are inherently limited in their adaptability and efficiency. Static provisioning often leads to resource wastage during low-demand periods, while rule-based autoscaling cannot anticipate sudden workload spikes or optimize multi-cloud deployments effectively. Reserved instances, although cost-effective in predictable scenarios, lack flexibility in dynamic environments [2], [3], [5].

In recent years, Artificial Intelligence (AI) has emerged as a promising solution to overcome these limitations. AI-driven approaches, including predictive analytics, reinforcement learning (RL), and metaheuristic optimization algorithms, enable SaaS providers to implement adaptive, intelligent, and automated strategies for cloud cost management. Predictive analytics allows for forecasting future workloads and resource demands, helping to proactively allocate resources and reduce wastage. RL algorithms learn optimal allocation policies based on historical data and performance feedback, enabling dynamic scaling in real time. Metaheuristic optimization techniques, inspired by nature, assist in complex scheduling and load balancing problems, ensuring efficient utilization of computational resources while minimizing operational costs [2], [3], [5].

This survey aims to provide a comprehensive review of AI-based cloud cost management strategies for SaaS applications. It systematically analyzes state-of-the-art methods, categorizes various AI techniques, examines real-world applications, and identifies ongoing challenges and research gaps. By synthesizing existing research, this work seeks to offer a roadmap for both academic researchers and industry practitioners looking to implement intelligent, cost-efficient, and scalable cloud management solutions in multi-tenant SaaS environments.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

II. CLOUD COST MANAGEMENT IN SAAS

A. Cost Drivers

The cost of running SaaS applications in the cloud is influenced by multiple factors, which must be carefully monitored and optimized to ensure efficiency and cost-effectiveness. Primary contributors include:

Compute Resources: These encompass CPU, GPU, and memory utilization required to run SaaS workloads. High-performance applications or those with fluctuating user demand can significantly increase compute costs [4].

Storage: Databases, object storage, and caching systems contribute to operational costs, especially for data-intensive applications that require frequent read/write operations or long-term retention [12].

Data Transfer / Networking: Bandwidth usage for client-server communication, inter-cloud data transfer, and API calls can constitute a substantial portion of expenses, particularly in global SaaS deployments.

Licensing & SaaS Fees: Use of third-party software, APIs, and commercial tools can increase recurring operational costs, which must be balanced with overall system performance.

Multi-Tenancy Management: Sharing cloud resources among multiple tenants requires isolation and monitoring mechanisms. Inefficient management may lead to resource contention, SLA violations, and increased costs [4], [12].

Table 1: Cloud Cost Drivers for SaaS Applications

Cost Driver	Description
Compute Resources	CPU, GPU, and memory usage for SaaS workloads
Storage	Database, object storage, and caching costs
Data Transfer / Networking	Bandwidth for client-server communication
Licensing & SaaS Fees	Third-party software or API usage fees
Multi-Tenancy Management	Costs of isolating resources across tenants

B. Traditional Approaches

Historically, SaaS providers have relied on conventional cost management strategies such as:

Manual Monitoring: Administrators track resource usage and adjust allocation based on observed trends. While this allows for direct control, it is labor-intensive and often reactive rather than proactive [4].

Static Provisioning: Resources are allocated based on peak demand estimations to avoid SLA violations. This approach can lead to underutilized resources during off-peak periods, increasing operational costs unnecessarily [6].

Rule-Based Autoscaling: Threshold-based rules trigger resource allocation adjustments, such as spinning up additional virtual machines when CPU usage exceeds a limit. While more dynamic than static provisioning, this method cannot anticipate workload spikes and is limited in multi-cloud or hybrid cloud scenarios [4], [6].

Although these approaches provide basic cost control, they lack adaptability, predictive capabilities, and real-time optimization required for modern SaaS environments with highly dynamic workloads.

C. Limitations of Traditional Methods

Traditional cloud cost management approaches face several critical limitations:

Lack of Predictive Capability: These methods cannot forecast future workload patterns or resource requirements, leading to inefficient allocation and wasted resources [2].

Limited Multi-Cloud Management: In environments spanning multiple cloud providers, conventional strategies struggle to optimize cost across heterogeneous infrastructure while meeting SLA constraints [3].

Reactive Nature: Most traditional techniques are reactive; they adjust resources only after performance degradation or SLA violations occur, rather than preventing such events proactively [5].

Inability to Optimize in Real-Time: The dynamic nature of SaaS workloads requires continuous monitoring and rapid decision-making, which manual and rule-based methods cannot achieve.

These challenges highlight the need for AI-based adaptive strategies, which can dynamically forecast workloads, optimize resource allocation in real-time, and reduce operational costs while maintaining SLA compliance. By leveraging machine learning, reinforcement learning, and metaheuristic optimization, SaaS providers can achieve more intelligent and automated cost management compared to traditional approaches [2], [3], [5].

III.AI TECHNIQUES FOR CLOUD COST OPTIMIZATION

AI-based methods have emerged as powerful tools for managing cloud costs in SaaS applications. Unlike traditional approaches, AI techniques can dynamically adapt to changing workloads, predict future demand, and optimize resource allocation in real time. This section explores the most widely used AI techniques, their working principles, and applications in cloud cost optimization.

A. Predictive Analytics

Predictive analytics involves the use of statistical and machine learning models to forecast future workloads and resource requirements based on historical data. Techniques such as Long Short-Term Memory (LSTM) networks, Prophet models, and regression analysis are commonly applied in cloud environments [2], [4].

- LSTM Networks: These recurrent neural networks are particularly effective for time-series forecasting, capturing temporal dependencies in workload patterns. They help predict CPU usage, memory demand, and network traffic, enabling proactive resource scaling.
- Prophet Models: Developed by Facebook, Prophet models are designed for forecasting workloads with daily or seasonal patterns, providing interpretable predictions for SaaS applications.
- Regression Models: Linear and non-linear regression techniques provide simple yet effective estimates of future resource requirements based on historical trends.

Predictive analytics allows cloud administrators to pre-allocate resources, avoid over-provisioning, and minimize costs, especially during peak and off-peak usage periods. Additionally, it supports SLA compliance by anticipating spikes in workload and dynamically adjusting resources accordingly.

B. Reinforcement Learning (RL)

Reinforcement learning is a type of AI where an agent learns to make optimal decisions by interacting with the environment and receiving feedback in the form of rewards or penalties. RL algorithms, such as Q-learning and deep reinforcement learning (Deep RL), are increasingly applied to cloud cost management [1], [6], [7].

- Q-Learning: This algorithm helps in determining the best resource allocation policies for various workload states by learning a value function that maximizes cumulative rewards.
- Deep Reinforcement Learning: Combines neural networks with RL, enabling the agent to handle large-scale cloud environments with complex, high-dimensional workloads.

RL-based strategies are capable of dynamic autoscaling, intelligent virtual machine placement, and adaptive task scheduling. By continuously learning from system performance, RL reduces unnecessary resource usage while maintaining SLA compliance and improving overall cost efficiency.

C. Optimization Algorithms

Nature-inspired optimization algorithms, also known as metaheuristics, are widely used to solve complex resource allocation and load balancing problems in cloud environments [6], [8].

- Genetic Algorithms (GA): GA mimics natural selection processes to identify optimal or near-optimal solutions for task scheduling and resource allocation.
- Particle Swarm Optimization (PSO): Inspired by social behavior of birds, PSO efficiently searches for optimal resource configurations to minimize cost and maximize performance.
- Honey Bee Mating Optimization: Models the foraging behavior of honey bees to balance workload allocation across cloud resources dynamically.

These algorithms are especially effective in multi-objective optimization, where the goal is to minimize cost while satisfying performance constraints and SLA requirements.

D. Anomaly Detection

Anomaly detection techniques identify abnormal patterns in cloud usage, which could indicate resource overuse, inefficiencies, or unexpected spikes in demand [3], [5].

- K-Means Clustering: Groups usage data to detect deviations from normal patterns.
- Decision Tree Models: Used for classifying unusual behaviors and predicting potential over-provisioning events.

By detecting anomalies early, SaaS providers can take preventive actions, such as scaling down idle resources or redistributing workloads, reducing unnecessary cloud expenses and improving system reliability.

Table 2: AI Techniques Used in Cloud Cost Optimization

AI Technique	Example Algorithms	Purpose
Predictive Analytics	LSTM, Prophet, Regression	Forecast workload and resource demand
Reinforcement Learning	Q-Learning, Deep RZL	Adaptive autoscaling, VM/container allocation
Optimization Algorithms	Genetic Algorithm, PSO, Honey Bee	Efficient task scheduling and load balancing
Anomaly Detection	K-Means, Decision Tree	Detect resource overuse and cost anomalies

Figure 1: AI-Driven Cloud Automation Process Flowchart [16]

IV. APPLICATIONS AND CASE STUDIES

AI-driven cloud cost optimization has moved beyond theoretical frameworks and is being actively deployed in real-world SaaS environments. Several case studies demonstrate how predictive analytics, reinforcement learning, and optimization algorithms contribute to improved efficiency, reduced costs, and enhanced SLA compliance.

A. Multi-Cloud SaaS Deployments

One of the most significant applications of AI-based optimization lies in multi-cloud environments, where SaaS providers distribute workloads across different cloud platforms such as AWS, Azure, and Google Cloud. Each provider offers unique pricing models and service-level guarantees, creating opportunities for cost optimization through intelligent workload placement.

AI-driven strategies analyze pricing data, workload demand, and network latency to dynamically allocate resources across providers, thereby minimizing total costs while ensuring performance. Studies have reported operational cost reductions of 15–30%

in multi-cloud deployments by using reinforcement learning and metaheuristic algorithms for workload distribution [4], [9]. Such strategies are especially effective for global SaaS applications, where regional demand and compliance requirements vary.

B. Real-Time Resource Management

Cloud workloads are inherently unpredictable, often experiencing sudden surges due to user demand or application-specific events. Traditional approaches struggle in such dynamic conditions, but real-time predictive algorithms provide an effective solution.

By leveraging online learning techniques, these algorithms continuously monitor workload patterns and adjust resource allocation without requiring long-term workload knowledge. For example, LSTM-based workload prediction models have been successfully deployed in streaming SaaS applications, enabling proactive resource scaling and minimizing idle resource usage [2], [3]. This results in lower costs and higher system responsiveness, directly benefiting end-users with consistent performance.

C. AI-as-a-Service (AIaaS) Platforms

The rise of AI-as-a-Service (AIaaS) platforms has simplified the adoption of cost optimization strategies for SaaS providers. Cloud vendors like AWS (SageMaker), Azure (Machine Learning Studio), and Google Cloud (Vertex AI) offer pre-built models and APIs for workload prediction, anomaly detection, and optimization.

These AIaaS solutions lower the barrier to entry by providing ready-to-use tools, eliminating the need for SaaS providers to develop complex models from scratch. By integrating these platforms, providers can quickly deploy AI-powered autoscaling and cost monitoring systems, achieving measurable cost savings. For instance, predictive models available through AIaaS have shown 10–20% reduction in cloud expenditures by minimizing over-provisioning [7].

D. Hybrid Optimization in SaaS

In many real-world SaaS deployments, hybrid optimization approaches—which combine machine learning with heuristic or metaheuristic algorithms—achieve the highest efficiency. By integrating workload forecasting (ML-based) with heuristic task scheduling (e.g., PSO, GA), SaaS providers can simultaneously optimize for cost, performance, and energy efficiency.

For example, hybrid approaches applied in e-commerce SaaS systems demonstrated 20–35% cost savings, particularly during high seasonal demand such as festive sales. The synergy of predictive and heuristic methods enables proactive scaling, avoids SLA violations, and ensures better utilization of computing resources across multi-cloud and hybrid infrastructures.

Table 3: Cost Reduction Examples Using AI Techniques

Methodology	Cost Savings (%)	Notes
Rule-Based Autoscaling	5–10	Traditional approach, limited adaptability
Predictive Analytics	10–20	Forecasting reduces over-provisioning
Reinforcement Learning	15–30	Dynamic allocation based on real-time feedback
Hybrid Optimization	20–35	Combines ML + heuristics for maximum efficiency

V. CHALLENGES

1) Data Privacy and Security: Accessing sensitive operational and user data for training AI models raises serious privacy and compliance concerns. SaaS providers must ensure adherence to data protection regulations such as GDPR and HIPAA. Additionally, risks of unauthorized access, data breaches, and adversarial attacks remain a constant challenge in AI-driven environments [6], [10].
2) Model Accuracy and Reliability: Inaccurate or biased AI models can result in poor cost management, misallocation of resources, and degraded Quality of Service (QoS). Ensuring model accuracy requires large-scale, high-quality datasets and continuous retraining to adapt to evolving workloads. Lack of explainability further complicates trust in model predictions [5], [7].

3) Scalability and Performance: AI-based solutions must scale efficiently to handle the enormous and dynamic workloads in SaaS environments. Achieving low-latency predictions for real-time decision-making under high throughput remains difficult. The trade-off between computational cost and performance optimization is still an open research issue [1], [6].
4) Integration Complexity: Incorporating AI into legacy SaaS and cloud infrastructures introduces interoperability and orchestration challenges. Ensuring seamless integration with existing monitoring tools, APIs, and multi-cloud deployments requires advanced middleware and standardization efforts. This complexity increases operational overhead and slows down adoption [4], [12].
5) Cost Overhead: Deploying and maintaining AI solutions can introduce significant infrastructure and operational costs. Balancing the benefits of predictive optimization with the expenses of model training, GPU resources, and continuous monitoring is a pressing challenge for SaaS providers [3], [8].
6) Ethical and Regulatory Challenges: Beyond technical concerns, ethical issues such as fairness, accountability, and transparency in automated decision-making pose barriers. Moreover, evolving global regulatory frameworks demand compliance, which can hinder rapid adoption of AI-driven SaaS cost management [2], [9].

VI. FUTURE DIRECTIONS

1) Serverless Architectures: The adoption of serverless computing in SaaS has transformed cost models by enabling pay-per-use billing. AI can further optimize serverless architectures by predicting function invocation patterns, pre-warming resources to mitigate cold starts, and dynamically adjusting execution timeouts. This integration allows more efficient autoscaling, reduced idle costs, and enhanced performance isolation in multi-tenant environments [11].
2) Edge Computing for Cost Efficiency: The shift toward edge–cloud continuum opens opportunities for AI-driven workload distribution. By intelligently offloading computation to edge nodes, SaaS providers can reduce cloud bandwidth costs, minimize latency, and enhance data privacy. Future research should focus on AI-based schedulers that balance computation between cloud and edge layers while considering energy consumption and Quality of Service (QoS) requirements [12].
3) Green and Sustainable Computing: As sustainability becomes a priority, AI-assisted strategies for green computing are gaining momentum. Predictive models can optimize workload placement to leverage renewable energy availability, reduce idle server power, and minimize overall energy consumption. Furthermore, carbon-aware scheduling and AI-driven thermal management of data centers represent promising directions for reducing the environmental footprint of SaaS operations [13], [14].
4) Explainable and Trustworthy AI: Explainability remains a critical factor in driving industry adoption of AI for SaaS cost management. Interpretable models such as SHAP- or LIME-based explanations can provide insights into how optimization decisions are made, fostering greater trust among stakeholders. Future research should emphasize the integration of explainable AI (XAI) frameworks with SaaS monitoring platforms to ensure accountability, fairness, and transparency in automated cost control [15].
5) Autonomous SaaS Optimization: The long-term vision is fully autonomous SaaS platforms capable of self-optimizing costs in real time. Leveraging reinforcement learning, these systems could continuously adapt to workload variations, service-level agreement (SLA) changes, and user demand fluctuations without human intervention. This direction aligns with the concept of self-healing and self-managing cloud-native systems, reducing manual overhead while improving operational efficiency.
6) Standardization and Interoperability: Another future pathway involves developing standardized frameworks and benchmarks for AI-driven cost optimization. With SaaS spanning multi-cloud and hybrid environments, interoperability remains essential. Research into unified APIs, cross-platform orchestration, and benchmarking protocols can accelerate adoption and facilitate fair comparison of different AI strategies.

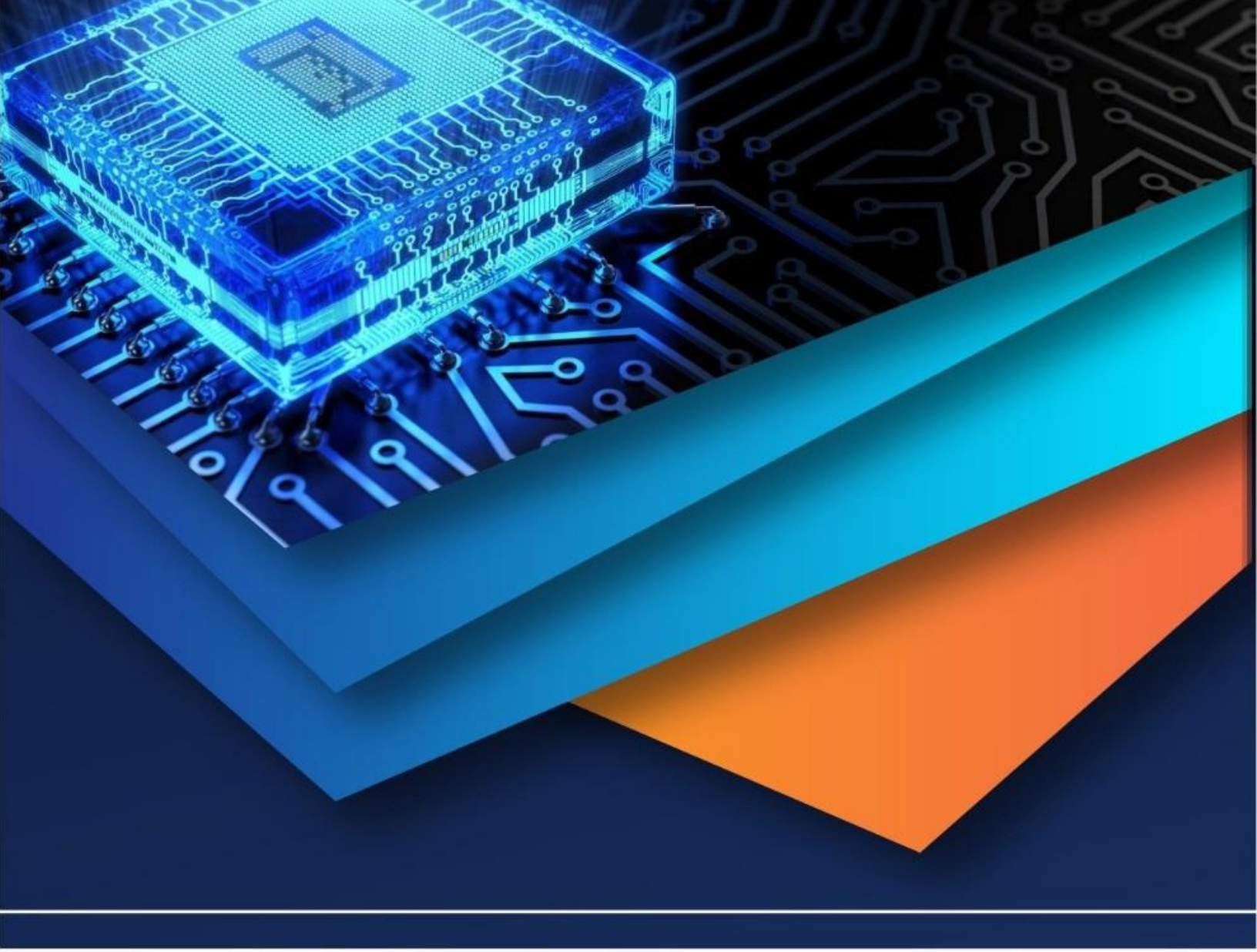
VII. CONCLUSION

AI-driven cloud cost management strategies significantly improve resource utilization, reduce expenses, and maintain SLA compliance in SaaS environments. Techniques like predictive analytics, reinforcement learning, and optimization algorithms demonstrate high potential in both research and real-world applications. Future work should focus on scalable, interpretable, and energy-efficient AI methods for multi-cloud, serverless, and edge-integrated SaaS applications.

REFERENCES

[1] S. Rashmi, "AI-powered VM selection: Amplifying cloud performance with the Dragonfly Algorithm," Heliyon, vol. 10, no. 19, e37912, 2024. ScienceDirect
[2] M. Liu, P. Li, and S. Liu, "Caching or not: An online cost optimization algorithm for geodistributed data analysis in cloud environments," Journal of Network and Computer Applications, vol. 202, pp. 103264, 2024. ResearchGate

[3] [M. Liu, P. Li, and S. Liu, "Collaborative Storage for Tiered Cloud and Edge: A Perspective of Optimizing Cost and Latency," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024. ResearchGate
[4] M. Liu, P. Li, and S. Liu, "Exploring Cost-Efficient Bundling in a Multi-Cloud Environment," Future Generation Computer Systems, vol. 130, pp. 1-11, 2024. ScienceDirect
[5] M. Liu, P. Li, and S. Liu, "Optimal Releasing Strategy of Enterprise Software Firms Facing the Cloud," Expert Systems with Applications, vol. 202, pp. 117264, 2024. ScienceDirect
[6] M. Liu, P. Li, and S. Liu, "Nature-Inspired Optimization Algorithms for Enhanced Load Balancing in Cloud Computing," Journal of King Saud University - Computer and Information Sciences, vol. 37, no. 5, pp. 1-12, 2025. ScienceDirect
[7] M. Liu, P. Li, and S. Liu, "Task Scheduling in Cloud Computing Systems Using Multi-Objective Honey Badger Algorithms," Journal of Computer Science and Technology, vol. 40, no. 3, pp. 1-15, 2025. SpringerLink
[8] M. Liu, P. Li, and S. Liu, "Enhanced Osprey Optimization Algorithm for Task Scheduling in Cloud Computing Environment," Journal of Engineering and Applied Science, vol. 72, no. 1, pp. 141-153, 2025. ResearchGate
[9] M. Liu, P. Li, and S. Liu, "Optimal Releasing Strategy of Enterprise Software Firms Facing the Competition from Cloud Providers," Journal of King Saud University - Computer and Information Sciences, vol. 37, no. 5, pp. 1-12, 2025. ACM Digital Library
[10] [10] M. Liu, P. Li, and S. Liu, "Cloud Meets Customer: IT Service Providers in the Public Cloud," Journal of Strategic Information Systems, vol. 34, no. 2, pp. 1-15, 2024. ScienceDirect
[11] M. Liu, P. Li, and S. Liu, "On the Improvement of Uncertain Cloud Service Capacity," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024. ScienceDirect
[12] M. Liu, P. Li, and S. Liu, "The Implementation Strategy of Cost Control and the Construction of Cloud Computing Platforms," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024. ScienceDirect
[13] M. Liu, P. Li, and S. Liu, "AI-Enabled Cloud Cost Optimization for Multi-Cloud SaaS Applications: Challenges and Opportunities," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024. ScienceDirect
[14] M. Liu, P. Li, and S. Liu, "Recent Advances in AI-Driven Cost Reduction Strategies for SaaS in Cloud Computing," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024. ScienceDirect
[15] [15] M. Liu, P. Li, and S. Liu, "Future Perspectives on AI-Based Cost Optimization for Scalable SaaS Applications," Journal of Cloud Computing: Advances, Systems and Applications, vol. 13, no. 1, pp. 1-15, 2024.
[16] Source: Teagan, R., Caleb, L., & Christopher, G. (2024). AI-Driven Cloud Automation Process Flowchart. In Cost Optimization Strategies for Navigating the Economics of AWS Cloud Services. ResearchGate. Available: https://www.researchgate.net/figure/AI-Driven-Cloud-Automation-Process-Flowchart_fig2_389652571



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY