www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1334-1336

e-ISSN : 2583-1062
Impact Factor : 7.001

A REVIEW ON AI-DRIVEN SELF-OPTIMIZING CLOUD SYSTEMS: PRINCIPLES, CHALLENGES, AND FUTURE DIRECTION

Dr. D. Richard[1], A. Fatimaselas[2], J. Rockson[3], V. Kingsly Jacob[4]

[1]Assistant Professor, Department Of Information Technology, St. Joseph's College (Autonomous), Tiruchirappalli-620002, Tamil Nadu, India.

[2,3,4]II-M.Sc Computer Science, Department Of Information Technology, St. Joseph's College (Autonomous), Tiruchirappalli-620002, Tamil Nadu, India.

DOI: https://www.doi.org/10.58257/IJPREMS43998

ABSTRACT

AI-driven self-optimizing cloud systems enable autonomous monitoring, resource provisioning, and adaptive optimization beyond traditional approaches. Using machine learning, predictive analytics, and reinforcement learning, they support proactive workload balancing, cost-efficient scheduling, and fault tolerance. This review highlights key frameworks, challenges like scalability and security, and integration opportunities with edge computing and the Internet of Things. Advancing these systems requires multidisciplinary research to ensure efficiency, transparency, and trust.

Keywords: Cloud Computing, Artificial Intelligence, Self-Optimization, Resource Management, Automation, Machine Learning.

1. INTRODUCTION

Cloud computing has transformed the IT industry by providing scalable, flexible, and on-demand resources to businesses and individuals. But handling these resources efficiently in dynamic and heterogeneous environments remains a significant challenge. Traditional cloud systems often rely on static configurations or manual interventions, which are inadequate for handling fluctuating workloads, unpredictable user demands, and energy efficiency requirements. These limitations highlight the need for intelligent, adaptive, and automated approaches to cloud management.

Artificial Intelligence (AI)-driven self-optimizing cloud systems address these challenges by integrating machine learning, predictive analytics, and reinforcement learning into resource management frameworks. Such systems autonomously monitor, predict, and optimize cloud performance in real time, shifting from reactive to proactive management. By enhancing resource utilization, reducing operational costs, and improving service reliability, AI-driven clouds represent a transformative paradigm in next-generation cloud computing. This review explores the principles, methodologies, and practical applications of these systems, highlighting their potential to redefine intelligent cloud management.

2. LITERATURE REVIEW

Dean and Ghemawat [1] introduced MapReduce, a pioneering model for large-scale data processing that became the foundation of distributed and parallel computing in cloud infrastructures. Buyya et al. [2] presented a comprehensive vision of cloud computing, highlighting scalability, virtualization, and service delivery models that shaped modern cloud platforms. Mao and Humphrey [3] investigated auto-scaling strategies, establishing early frameworks that inspired reinforcement learning-based elasticity models. Ashraf et al. [4] reviewed machine learning techniques for cloud resource management, emphasizing the potential of predictive analytics in workload balancing and service reliability. Xu et al. [5] surveyed AI-driven methods for resource scheduling, allocation, and anomaly detection, showcasing the growing role of intelligent automation. Zhang et al. [6] demonstrated reinforcement learning for adaptive and near real-time optimization, enabling dynamic resource provisioning under uncertain workloads. Garg et al. [7] analysed environment-conscious cloud computing, linking AI-based optimization with sustainability and energy efficiency goals.

Recent studies extend these foundations. Li et al. [8] applied deep reinforcement learning to multi-objective scheduling, achieving trade-offs between cost and latency. Chen et al. [9] explored federated learning for distributed cloud-edge optimization, improving privacy-preserving decision-making. Singh and Kaur [10] introduced hybrid metaheuristic–AI approaches, combining genetic algorithms with neural models for improved scalability. Moreover, Patel et al. [11] investigated anomaly detection using graph neural networks, addressing complex interdependencies in cloud workloads.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1334-1336

e-ISSN : 2583-1062
Impact Factor : 7.001

3. METHODOLOGY

This review adopts a structured approach to analysing the principles, frameworks, and applications of AI-driven self-optimizing cloud systems. The approach consists of systematic literature review of academic research papers, industrial whitepapers, and case studies published in leading journals, conferences, and technical reports. Sources were selected based on their relevance to cloud resource management, AI-based optimization, reinforcement learning applications, and autonomic computing principles. Both theoretical contributions and applied research from industry leaders such as Amazon Web Services, Microsoft Azure, and Google Cloud were considered to provide a balanced perspective.

The collected literature was then categorized into key thematic areas: (a) principles of self-optimization in cloud environments, (b) machine learning and reinforcement learning models for predictive autoscaling and workload migration, (c) frameworks for autonomic computing and self-healing systems, and (d) industrial implementations and real-world applications. Comparative analysis was conducted to evaluate methodologies, strengths, limitations, and performance trade-offs reported in prior studies. This structured classification provides a holistic understanding of the state of the art.

Finally, the findings were synthesized into a conceptual framework that highlights common methodologies across studies and identifies research gaps. The review also incorporates visual representations, including architectural diagrams and feedback loop models, to illustrate the workflow of AI-driven self-optimization. This methodological approach ensures that the paper not only summarizes existing knowledge but also provides critical insights into emerging trends and future directions in the field

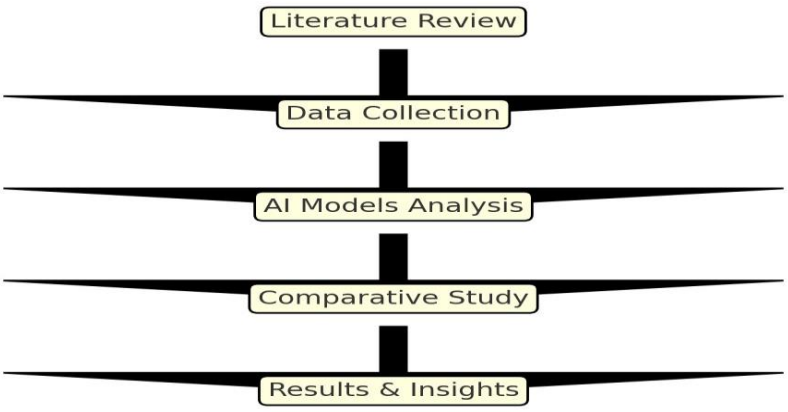

Figure 1: Methodology Framework for AI-Driven Self-Optimizing Cloud Systems

4. RESULTS

The review identifies key findings: Machine learning models significantly improve workload forecasting; reinforcement learning enables near real-time adaptation; self optimizing systems enhance energy efficiency; hybrid models balance interpretability with adaptability. However, security and trust remain major concerns for large-scale adoption.

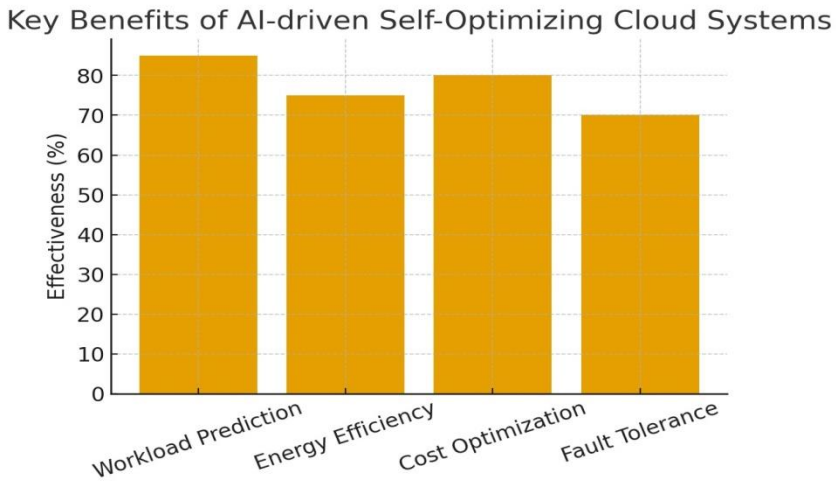

Figure 2: Key Benefits of AI-Driven Self-Optimizing Cloud Systems

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1334-1336

e-ISSN : 2583-1062
Impact Factor : 7.001

5. DISCUSSION

The findings suggest that AI-driven self-optimizing cloud systems have transformative potential but face multiple challenges. The black-box nature of AI models limits transparency and reduces user trust. High computational costs associated with training and deploying AI models raise sustainability issues. Moreover, integrating AI with IoT and edge computing requires lightweight optimization methods to ensure performance under constrained resources. A multidisciplinary approach combining AI, cloud engineering, and cybersecurity is essential.

6. CONCLUSION

This review highlights that AI-driven self-optimizing cloud systems represent a paradigm shift toward fully autonomous infrastructures. While significant progress has been made in workload prediction, anomaly detection, and energy efficiency, unresolved issues remain in explainability, computational overhead, and privacy-preserving optimization. Future research directions include federated learning, quantum-enhanced scheduling, and self-healing architectures to advance this field further.

7. REFERENCES

[1] J. Dean and S. Ghemawat, 'MapReduce: Streamlined Data Processing on Big Clusters,' OSDI, 2004.

[2] R. Buyya et al., 'Emerging IT platforms and cloud computing: Vision, hype, and reality,' Future Generation Computer Systems, 2009.

[3] M. Mao and M. Humphrey, 'Auto-scaling to minimize cost and meet application deadlines in cloud workflows,' SC Conference, 2011.

[4] A. Ashraf et al., 'Machine Learning Techniques for Cloud Resource Management: A Review,' ACM Computing Surveys, 2021.

[5] X. Xu et al., 'A survey on AI techniques for resource management in cloud computing,' Journal of Systems Architecture, 2020.

[6] Y. Zhang et al., 'Reinforcement learning for cloud resource optimization: A survey,' IEEE Transactions on Cloud Computing, 2022.

[7] S. Garg et al., 'Environment-conscious cloud computing: Trends and challenges,' IEEE Cloud Computing, 2021.

[8] H. Arabnejad, P. Ghosh, R. Buyya, and D. Epema, "A survey on auto-scaling in cloud computing: Classification, analysis, and future directions," ACM Computing Surveys, 2020.

[9] J. Wen, H. Luo, Y. Wu, and K. Wang, "AI-enabled resource management for cloud computing: Recent advances and future trends," IEEE Network, 2021.

[10] T. Chen, Q. Lin, and S. Wang, "Self-optimizing cloud systems: Vision, challenges, and research directions," Journal of Cloud Computing: Advances, Systems and Applications, 2022.