# A Review of AI-Driven Techniques for Cost Optimization in Kubernetes Environments

Sivagnanasothy Lakshan[1*], Sahdiya Hussain[2]

[1]*Informatics Institute of Technology, Sri Lanka, lakshan.20232392@iit.ac.lk*
[2]*Informatics Institute of Technology, Sri Lanka, sahdiya.h@iit.ac.lk*

Kubernetes plays a pivotal role in managing cloud-native applications by automating deployment, scaling, and resource allocation. Despite its robust capabilities, it faces significant challenges in cost optimization due to over-provisioning, under-utilization, and the inability to effectively predict workload variations. Native Kubernetes autoscaling mechanisms such as Horizontal Pod Autoscaler (HPA), Vertical Pod Autoscaler (VPA) and Cluster Autoscaler (CA) often fail to strike an optimal balance between resource utilization and application performance. To address these limitations, recent researches has focused on leveraging AI and machine learning (ML) techniques to enhance Kubernetes' resource allocation efficiency. This review comprehensively examines the state-of-the-art advancements in AI-driven cost optimization strategies for Kubernetes environments. It explores predictive scaling algorithms, reinforcement learning-based orchestration, time-series forecasting models and clustering methods aimed at improving resource utilization and reducing operational expenses. The integration of ML technologies has demonstrated remarkable success in enabling proactive and dynamic scaling decisions, enhancing cost-effectiveness while maintaining high application reliability. Key studies highlight the potential of hybrid ML models, multi-objective optimization approaches and advanced metrics-driven strategies to address the complexities of resource allocation. These methodologies not only minimize inefficiencies but also provide adaptive solutions to evolving workload demands. Furthermore, the paper identifies gaps in existing approaches, such as the limited incorporation of multi-dimensional metrics and the need for continuous model adaptability. This review underscores the transformative potential of AI-driven techniques in Kubernetes environments, presenting a roadmap for future research to refine and expand these strategies. The findings aim to guide organizations in adopting intelligent autoscaling mechanisms to optimize costs, improve resource utilization and drive innovation in cloud-native application management.

**Keywords:** *AI, Cloud computing, Cost optimization, Kubernetes*

**Paper No: SE-03** **Software Engineering**