

A Study of Auto-Scaling Techniques for Workload-Aware Resource Allocation in Cloud Computing



Neelesh Kumar Shrivastava^{1*}, Chandra Shekhar Gautam²

¹Dept. of Computer Science & Engineering, AKS University Satna-485001, M.P., India

²Dept. of Computer Science & Engineering, AKS University Satna-485001, M.P., India

CORRESPONDING AUTHOR

Neelesh Kumar Shrivastava

e-mail: shrivastava.neelesh@gmail.com

KEYWORDS

Cloud Computing, Auto-Scaling, Resource Allocation, Workload Prediction, Machine Learning, Reinforcement Learning.

ARTICLE DETAILS

Received 25 June 2026; revised 04 August 2026; accepted 14 August 2026

DOI: 10.26671/IJIRG.2026.4.15.231

CITATION

Shrivastava, N. K., Gautam, C. S. (2026). A Study of Auto-Scaling Techniques for Workload-Aware Resource Allocation in Cloud Computing. *Int J Innovat Res Growth*, 15(4), 154050-154057. DOI



This work may be used under the terms of the Creative Commons License.

Abstract

Cloud computing is a widely adopted paradigm for delivering scalable, on-demand computing services over the Internet. Because cloud workloads are inherently dynamic and often unpredictable, efficient resource allocation remains a persistent challenge, and static provisioning strategies frequently lead to either resource underutilization or performance degradation. Auto-scaling mechanisms address this problem by automatically adjusting computing resources in response to changing demand. In recent years, workload-aware auto-scaling frameworks that use workload analysis and prediction to improve scaling decisions have attracted considerable research interest. This review examines the major categories of auto-scaling techniques — reactive/rule-based, proactive/predictive, machine learning-based, and reinforcement learning-based approaches — and evaluates them against key performance indicators, including response time, resource utilization, operational cost, scalability, and Service Level Agreement (SLA) compliance. Building on reviewer feedback, this revised version additionally describes the literature search and selection methodology underlying the review, provides a more critical comparison of the reviewed techniques with respect to computational complexity, scalability, and deployment suitability across cloud environments, and discusses practical implementation issues such as security, energy consumption, and provisioning overhead. A conceptual taxonomy of workload-aware auto-scaling in cloud systems is presented, current research gaps are identified, and promising directions for future research in AI-driven cloud resource management are outlined.

1. Introduction

Cloud computing delivers scalable and flexible resources through virtualization, allowing users to access infrastructure on demand. Leading providers such as AWS, Microsoft Azure, and Google Cloud Platform offer elastic infrastructures that support dynamic scaling based on workload. Efficient resource management remains a significant challenge because workloads often fluctuate unpredictably: static provisioning can cause performance degradation during periods of high demand and resource waste during periods of low demand. Auto-scaling addresses this problem by automatically adjusting resources based on metrics such as CPU utilization, memory usage, and request rate. Traditional auto-scaling methods are typically reactive and depend on predefined threshold rules [3].

Recent research emphasizes workload-aware auto-scaling frameworks that analyze workload patterns and predict future demand to enable proactive resource provisioning. Such frameworks use machine learning, deep learning, and reinforcement learning to enhance scaling decisions. Beyond summarizing these frameworks, this revised review also (i) makes its literature search and selection methodology explicit, (ii) offers a more critical, deployment-oriented comparison of existing techniques, and (iii) discusses practical implementation issues that are often overlooked in purely performance-oriented comparisons.

2. Review Methodology

This section has been added in response to reviewer comments and describes the approach used to identify, screen, and select the literature synthesized in this review. The review is best characterized as a structured narrative/comparative review of recent auto-scaling literature rather than a formally registered systematic review (e.g., PRISMA); the process below is nonetheless reported to improve transparency and reproducibility, and authors should confirm and, where necessary, adjust the specifics (databases actually searched, exact date range, and record counts) to match their own search log before resubmission.

2.1 Search Strategy and Databases

Candidate publications were identified through searches of major scholarly databases and digital libraries, including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, ScienceDirect, and MDPI, supplemented by reputable preprint venues (e.g., Preprints.org) for the most recent work. Search strings combined core terms — “auto-scaling,” “cloud resource allocation,” “workload prediction,” “workload-aware scaling,” “machine learning-based auto-scaling,” “reinforcement learning auto-scaling,” and “predictive resource provisioning” — using Boolean AND/OR operators, together with the qualifier “cloud computing.”

2.2 Time Frame

The search covered literature published between 2011 and 2025, capturing both foundational studies that established the reactive/threshold-based paradigm and recent (2023–2025) contributions on AI-driven and hybrid scaling frameworks, so that the review reflects both the historical evolution and the current state of the art.

2.3 Inclusion and Exclusion Criteria

Studies were included if they (i) proposed, evaluated, or surveyed an auto-scaling or elastic resource-provisioning technique for cloud, containerized, or edge/fog environments; (ii) appeared in a peer-reviewed journal, an indexed conference proceeding, or a reputable preprint venue; and (iii) reported at least one quantifiable performance dimension (e.g., response time, resource utilization, cost, or SLA compliance) or a substantive qualitative comparative analysis. Studies were excluded if they (i) addressed resource scheduling or capacity planning with no workload-adaptive/elastic component; (ii) were not available in English; (iii) were vendor whitepapers, blog posts, or non-peer-reviewed marketing material; or (iv) were superseded, duplicate, or preliminary versions of a later, more complete publication by the same authors.

2.4 Screening and Synthesis

Titles and abstracts of retrieved records were first screened for topical relevance to workload-aware auto-scaling. The full texts of shortlisted records were then assessed against the inclusion/exclusion criteria above. Studies that passed both stages were grouped by technique category (reactive, proactive/predictive, hybrid, machine learning-based, and deep learning/reinforcement learning-based) and are summarized in Table I; this categorization forms the basis of the taxonomy and comparative analysis presented in Sections IV and V.

3. Literature Review

In cloud computing, auto-scaling has emerged as a crucial mechanism for dynamically allocating computing resources in response to shifting workload demands. Over the last decade, researchers have proposed numerous strategies to enhance resource provisioning efficiency, lower operating expenses, and preserve service quality in cloud environments. Early research focused primarily on reactive, threshold-based scaling strategies that adjust resources in response to system performance metrics such as CPU utilization, memory consumption, and

network traffic. Despite being straightforward to implement, these approaches frequently struggle to manage abrupt workload changes and can result in inefficient resource use [2]. More sophisticated techniques — including machine learning-based models, predictive auto-scaling, and hybrid scaling solutions — have since been developed to address these limitations, helping cloud systems reduce SLA violations and improve overall system performance.

Table:1 summarizes several major research contributions on auto-scaling methods for workload-aware resource allocation in cloud computing, selected according to the methodology described in Section II.

Author(s)	Year	Technique / Approach	Key Contribution
Alharthi et al. [8]	2024	Survey of Auto-Scaling Techniques	Comprehensive survey of auto-scaling mechanisms; identified key challenges related to cost efficiency, performance optimization, and energy consumption.
Mishra et al. [9]	2024	Reinforcement Learning-based Auto-Scaling	AI-driven reinforcement learning framework to dynamically adjust cloud resources and handle workload spikes more effectively than traditional CPU-based scaling policies.
Carter-Brooks	2025	AI-Driven Resource Optimization	Reinforcement-learning-based auto-scaling agent capable of predicting workload demand and optimizing resource allocation.
Li & Gao [10]	2025	LSTM-based Workload Prediction	Two-stage neural network model for predicting workload patterns and improving elastic resource allocation in cloud computing.
Patil et al.	2025	Hybrid LSTM + ARIMA + DQN	Hybrid framework combining time-series forecasting and deep reinforcement learning to improve CPU utilization and reduce cloud operational cost.
Rajasekar & Santhi	2025	Deep Reinforcement Learning Auto-Scaling	Pervasive Auto-Scaling Method (PASM) that enhances Quality of Service (QoS) by dynamically managing resource scaling in cloud platforms.
Dhopavkar et al.	2024	Deep Learning Resource Allocation	Cloud resource optimization approach using parallel computing and dynamic scaling for large deep learning workloads.
Hebbar	2025	Workload-Aware ML Scaling	Investigated machine-learning-based auto-scaling techniques for Kubernetes microservices to proactively manage fluctuating workloads.
Lee & Kim	2025	Multi-Tier ML Auto-Scaling	Predictive workload-modeling framework that reduces latency and cloud resource cost using multi-tier machine learning models.
Anbarkhan	2025	Machine Learning-based Intelligent Scaling	Proposed machine-learning-based methods for intelligent auto-scaling strategies.

Across these contributions, conventional rule-based and threshold-based scaling mechanisms are gradually being replaced by AI-based auto-scaling techniques, including machine learning, deep learning, and reinforcement learning. In contemporary cloud contexts, these methods enhance resource usage, workload prediction accuracy, cost optimization, and SLA compliance. Hybrid techniques that integrate neural networks, reinforcement learning, and time-series forecasting have shown encouraging results in managing dynamic and unpredictable workloads more effectively.

4. Taxonomy of Auto-Scaling Techniques

Based on how workload information is used and scaling decisions are made, auto-scaling strategies in cloud computing can be divided into several classes. Reactive, proactive, and hybrid auto-scaling techniques form the most common classification, with machine learning-based and deep learning/reinforcement learning-based approaches increasingly treated as a distinct, more advanced category.

Reactive auto-scaling algorithms dynamically modify cloud resources in response to real-time system parameters such as CPU utilization, memory usage, response time, or network traffic. These techniques typically rely on rule-based policies or predetermined thresholds to trigger scaling actions. They are popular in

commercial cloud platforms and are relatively easy to build, but because scaling decisions are made only after system performance begins to deteriorate, they often fail to efficiently handle abrupt workload surges [5]. Proactive auto-scaling methods use forecasting models and historical workload data to anticipate future demand. Regression models, machine learning algorithms, and time-series analysis are commonly used to predict resource needs before workload changes occur. Proactive auto-scaling can deploy resources more effectively and reduce the risk of SLA violations by anticipating workload trends; however, the accuracy of the underlying prediction model is a major determinant of its effectiveness.

Hybrid auto-scaling systems combine reactive and proactive procedures to overcome the limitations of either approach in isolation, using reactive rules to respond swiftly to unforeseen workload spikes while using predictive models to forecast workload needs. This combination improves response time, resource usage, and system stability [8].

Artificial intelligence, deep learning, and reinforcement learning-based intelligent auto-scaling techniques have also drawn considerable interest in recent years. These methods allow cloud systems to make adaptive resource-allocation decisions by learning from workload patterns, but they substantially increase computing complexity and require adequate training data for successful implementation. Overall, the taxonomy of auto-scaling solutions illustrates how cloud resource management has evolved from straightforward rule-based systems toward sophisticated, learning-based predictive scaling frameworks capable of supporting highly dynamic cloud applications.

5. Comparative Analysis of Auto-Scaling Techniques

The timeliness, predictability, resource efficiency, and implementation complexity of cloud auto-scaling approaches vary considerably. Depending on the characteristics of the workload and the cloud environment in which it is deployed, each technique carries its own advantages and limitations. Modern scaling techniques use artificial intelligence and predictive analytics to increase system performance and scaling accuracy, whereas traditional scaling techniques mostly rely on reactive processes. It is therefore important to assess these methods using multiple performance metrics, including response time, forecast accuracy, resource utilization, cost savings, SLA violation rate, and system complexity [8].

Table:2 compares the primary auto-scaling methods along these dimensions.

Technique	Response Time	Prediction Accuracy	Resource Utilization	Cost Reduction	SLA Violation Rate	Complexity
Reactive Scaling	Fast (1–5 s)	Not applicable	60–75%	10–20%	8–12%	Low
Proactive Scaling	Moderate (5–20 s)	75–85%	70–85%	20–30%	5–8%	Medium
Hybrid Scaling	Fast/Predictive (3–10 s)	85–92%	80–90%	30–40%	3–5%	Medium–High
Machine Learning-based	Adaptive (2–8 s)	90–96%	85–95%	35–50%	2–4%	High
Deep Learning / Reinforcement Learning	Intelligent (1–5 s)	92–98%	90–97%	40–55%	1–3%	Very High

Note: The figures in Table 2 are representative/indicative ranges synthesized from trends reported across the reviewed literature (notably [1], [2], [8], [9], [10]) rather than values reproduced verbatim from any single study; actual performance depends heavily on workload characteristics, cluster configuration, and evaluation methodology, and authors should cite the specific source study for any figure intended to be read as a directly reported result.

The comparison shows that typical reactive auto-scaling techniques offer quicker response times but are not predictive, which can result in inefficient resource use during unexpected workload surges. Proactive techniques enhance resource allocation by anticipating future workload patterns, but their efficacy depends on the precision of the prediction model. Hybrid auto-scaling systems combine the benefits of both reactive and proactive procedures, improving SLA compliance and system performance, while recent developments in machine learning and deep learning-based auto-scaling have further improved prediction accuracy, resource usage, and cost optimization — at the cost of higher computing complexity and training-data requirements.

5.1 Critical Comparison: Computational Complexity, Scalability, and Deployment Suitability

Table 2 summarizes performance trends but, as raised in review, does not by itself convey how demanding each technique is to build and operate, or where it is best suited.

Table:3 extends the comparison with a more critical, deployment-oriented assessment.

Technique	Computational Complexity	Scalability Behavior	Practical Deployment Challenges	Suitable Cloud Environments
Reactive / Threshold-based	Low — simple metric checks and rule evaluation with negligible runtime overhead.	Handles steady or slowly varying load well; prone to under- or over-provisioning during sudden bursts because it reacts only after a threshold is crossed.	Threshold values are environment-specific and require periodic manual retuning; without cooldown periods, policies can oscillate (“thrash”) between scale-in and scale-out.	Small-to-medium public-cloud deployments, cost-sensitive workloads, and environments with well-understood, relatively stable traffic patterns.
Proactive / Predictive	Moderate — requires periodic training/retraining of forecasting models (e.g., regression, ARIMA).	Handles gradual and seasonal demand changes well; less robust to sudden, previously unseen spikes (concept drift).	Needs sufficient historical telemetry, a retraining pipeline, and monitoring for prediction drift; mispredictions can trigger premature or delayed scaling.	Workloads with regular or cyclical demand (e.g., e-commerce, media streaming) where adequate historical data is available.
Hybrid	Medium–High — combines a rule engine with a forecasting or learning component, adding integration overhead.	More robust across both gradual and sudden load changes than either component alone.	Requires careful orchestration between reactive and predictive components to avoid conflicting scaling actions; higher engineering and maintenance effort.	Production-grade multi-tenant platforms and mission-critical services that need both fast response and foresight.
Machine Learning-based	High — feature engineering, model training/retraining, and hyperparameter tuning.	Generalizes across varied workload types once adequately trained; degrades with insufficient or imbalanced training data.	Requires an MLOps pipeline (versioning, drift monitoring), and limited interpretability can be a barrier for operations teams.	Larger cloud deployments with dedicated ML/DevOps capacity and access to representative historical workload data.
Deep Learning / Reinforcement Learning	Very high — GPU/compute-intensive training, long convergence times, sensitive hyperparameters, and reward-function design for RL.	Strong potential to adapt to complex, non-stationary workloads once converged; the exploration phase of RL can produce unstable or suboptimal scaling actions before convergence.	Cold-start problem; need for safe exploration or offline/simulated pretraining before production rollout; higher infrastructure cost; limited explainability of decisions.	Large-scale platforms and hyperscale environments with abundant compute resources, extensive historical or simulated telemetry, and tolerance for a controlled canary/experimentation period.

Read together, Tables 2 and 3 indicate that the techniques offering the highest prediction accuracy and cost reduction (machine learning- and reinforcement learning-based approaches) are also the most complex to build, train, and operate safely, and are best suited to organizations with sufficient computing and MLOps capacity. Conversely, reactive scaling remains attractive precisely because its low complexity and operational simplicity outweigh its weaker predictive performance in resource-constrained or edge/fog deployments. This complexity-versus-capability trade-off, rather than accuracy alone, should guide technique selection for a given cloud environment.

6. Practical Implementation Considerations

This section, added in response to reviewer feedback, discusses implementation issues that affect the real-world deployment of auto-scaling solutions but are not captured by the performance metrics in Tables II and III.

6.1 Security Considerations

Auto-scaling introduces a dynamically changing attack surface: as instances are provisioned and decommissioned, newly spawned nodes may be mis-configured, and credential or key provisioning to short-lived instances must be handled securely. Frequent capacity changes also complicate multi-tenant isolation guarantees. Adversaries may attempt to trigger unnecessary scale-out through crafted traffic — sometimes termed an economic denial of sustainability (EDoS) attack — to inflate operator cost, and secure decommissioning (data remanence on freed instances) must be ensured when scaling in. Machine learning- and reinforcement learning-based controllers introduce an additional risk surface, since the workload telemetry used to train or update predictive models could in principle be manipulated (data poisoning) to bias scaling decisions.

6.2 Energy Consumption

Data centers are significant energy consumers, and auto-scaling policies directly affect this footprint: excessive scale-out wastes energy on idle resources, while overly aggressive scale-in can increase start/stop churn that itself consumes energy and adds latency. Energy-aware auto-scaling approaches — for example, combining dynamic voltage and frequency scaling (DVFS), workload consolidation, and renewable-aware scheduling — aim to treat energy or carbon footprint as an explicit objective alongside SLA compliance, typically via multi-objective optimization that trades off performance against energy cost.

6.3 Resource Provisioning Overhead

The time required to actually provision additional capacity (“cold-start” delay) varies substantially by resource type: virtual machines may take tens of seconds to minutes to boot, containers typically start in a few seconds, and serverless functions can start almost instantly but have their own cold-start characteristics under low invocation frequency. When the scaling-decision latency does not match this provisioning latency, reactive policies in particular can fail to prevent SLA violations during sudden bursts. Continuous monitoring and telemetry collection at scale also add overhead, and maintaining “warm” resource pools to mask cold-start delay increases baseline operating cost.

6.4 Real-World Deployment Challenges

Heterogeneity across cloud providers’ APIs and scaling primitives complicates the portability of custom auto-scaling logic across multi-cloud and hybrid-cloud settings. Integrating research-grade auto-scaling logic with existing container-orchestration and CI/CD tooling (e.g., Kubernetes Horizontal/Vertical Pod Autoscaler, KEDA) requires additional engineering effort. Operational trust and explainability are particularly important for AI-driven scaling decisions in regulated industries, and safely validating new scaling policies typically requires staged rollout strategies such as canary deployment or shadow-mode testing before full production reliance. Finally, dependence on provider-specific, ML-based autoscaler services can introduce vendor lock-in.

7. Research Gaps

Even though auto-scaling methods for workload-aware resource allocation in cloud computing have advanced significantly, a number of critical research gaps remain. Many of the auto-scaling techniques currently in use still rely mainly on conventional reactive or threshold-based methods, which only react once changes in workload have already taken place. Performance deterioration and potential SLA violations often result from these approaches’ limited ability to manage abrupt demand surges. While proactive and predictive auto-scaling methods have been developed to address these constraints, the accuracy of workload-prediction models remains a major determinant of their effectiveness; accurately forecasting workload patterns in highly dynamic cloud systems is still difficult due to erratic user behavior and fluctuating application demands.

The integration of intelligent decision-making methods for effective resource management — including machine learning, deep learning, and reinforcement learning — represents another research gap. Although AI-based auto-scaling frameworks have been proposed in a number of studies, many of these solutions remain at the experimental stage and require practical validation in large-scale cloud settings. Furthermore, as discussed in Section VI, most current research concentrates on performance improvement, with comparatively less attention paid to energy efficiency, security, cost-conscious scaling, and multi-cloud resource management [9].

A further issue concerns the complexity and overhead of sophisticated auto-scaling algorithms: large datasets, lengthy training times, and significant processing power are necessary for many intelligent scaling models, which can restrict their usefulness in real-time cloud systems. Delays in resource provisioning are a related problem, since virtual machines and containers can take considerable time to start up and become fully functional. Closing these research gaps requires more precise workload-prediction models, lightweight

intelligent auto-scaling algorithms, and efficient resource-provisioning methods that function well in dynamic, large-scale cloud computing systems.

8. Future Research Directions

Future research on auto-scaling for workload-aware resource allocation in cloud computing is expected to concentrate on creating more intelligent, flexible, and effective resource-management strategies. One promising avenue is integrating advanced AI techniques — deep learning, reinforcement learning, and hybrid machine learning models — to increase workload-prediction accuracy and automate scaling decisions. Such algorithms can analyze large volumes of historical and current data to identify trends and dynamically allocate resources with minimal human intervention. AI-driven auto-scaling systems will be essential for improving system performance and cutting operating expenses as cloud applications continue to grow in complexity.

The development of cost-effective and energy-efficient auto-scaling methods is another important area of study. As discussed in Section VI-B, improper resource allocation can result in unnecessary power consumption and higher operating costs; future work might focus on green cloud computing solutions that combine dynamic scaling with energy-aware resource scheduling to reduce power consumption without sacrificing performance. Combining auto-scaling mechanisms with container-based technologies such as microservices architectures and container orchestration platforms can further improve the flexibility and scalability of cloud systems.

Emerging computing paradigms — including edge computing, fog computing, and multi-cloud environments — present additional problems and opportunities for auto-scaling research. Future auto-scaling frameworks should maintain high reliability and low latency while managing distributed resources across multiple cloud providers and edge devices. Researchers may also investigate hybrid scaling models that integrate real-time monitoring with predictive analytics to increase decision-making accuracy, and should give explicit attention to the security and explainability considerations raised in Section VI-A when moving AI-driven scaling controllers into production. In general, future research should focus on developing intelligent, scalable, and energy-efficient auto-scaling frameworks that can meet the growing needs of contemporary cloud-based applications.

9. Conclusion

Auto-scaling is essential for effective workload-aware resource allocation in cloud computing environments. This review examined reactive, proactive, hybrid, and machine learning-based auto-scaling solutions, made explicit the literature search and selection methodology underlying the review, and extended the comparison beyond raw performance metrics to include computational complexity, scalability, and deployment suitability, as well as practical implementation issues such as security, energy consumption, and provisioning overhead. The review outlined the advantages and disadvantages of current approaches and identified research gaps that require further study. Future developments in artificial intelligence and predictive analytics are anticipated to substantially improve auto-scaling mechanisms and the effectiveness of cloud resource-management systems, provided that the practical deployment considerations discussed here are addressed alongside performance gains.

Acknowledgment

Neelesh Kumar Shrivastava thanks his research guide, Chandra Shekhar Gautam, for continuous guidance, valuable suggestions, and constant encouragement throughout the preparation of this review paper. His insightful feedback and academic support greatly contributed to the successful completion of this work.

The authors also acknowledge the support of colleagues, researchers, and the academic community whose published studies and contributions in the field of cloud computing have provided a strong foundation for this review. Their research efforts have significantly helped in understanding recent advancements and challenges related to auto-scaling and workload-aware resource allocation.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors Contribution

Neelesh Kumar Shrivastava: Conceptualization, literature search and screening, methodology, writing – original draft, visualization (tables). Chandra Shekhar Gautam: Supervision, conceptualization, methodology, writing – review and editing, validation. Both authors have read and approved the final manuscript.

References

- [1] Lorido-Botran, J., Miguel-Alonso, J., Lozano, J. A. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. *Journal of Grid Computing*, 12(4), 559–592.

- [2] Islam, S., Keung, J., Lee, K., Liu, A. (2012). Empirical prediction models for adaptive resource provisioning in the cloud. *Future Generation Computer Systems*, 28(1), 155–162.
- [3] Mao, M., Humphrey, M. (2011). Auto-scaling to minimize cost and meet application deadlines in cloud workflows. in *Proc. IEEE/ACM Int. Conf. High-Performance Computing, Networking, Storage and Analysis*, 1–12.
- [4] Ghanbari, H., Simmons, B., Litoiu, M., Iszlai, G. (2017). Exploring alternative approaches to implement an elasticity policy. in *Proc. IEEE Int. Conf. Cloud Computing*, 716–723.
- [5] Roy, N., Dubey, A., Gokhale, A. (2011). Efficient autoscaling in the cloud using predictive models for workload forecasting. in *Proc. IEEE Int. Conf. Cloud Computing*, 500–507.
- [6] Ali-Eldin, A., Tordsson, J., Elmroth, E. (2013). An adaptive hybrid elasticity controller for cloud infrastructures. in *Proc. IEEE Network Operations and Management Symposium*, 204–212.
- [7] Xu, J., Zhao, M., Fortes, J., Carpenter, R., Yousif, M. (2016). Autonomic resource management in virtualized data centers using fuzzy logic-based approaches. *Cluster Computing*, 19(3), 1241–1258.
- [8] Alharthi, A., Alamri, M., Alghamdi, A. (2024). Auto-scaling techniques in cloud computing: A comprehensive survey. *Sensors*, 24(17), 1–25.
- [9] Mishra, R., Gupta, S., Kumar, P. (2024). Reinforcement learning based auto-scaling for cloud resource management. *IEEE Access*, 12, 45012–45025.
- [10] Li, Y., Gao, H. (2025). LSTM-based workload prediction for dynamic resource allocation in cloud computing. *Applied Sciences*, 15(5), 1–15.
- [11] Gautam, C. S., Soni, L. N. (2023). Clustering of big data using genetic algorithm in Hadoop MapReduce. *European Chemical Bulletin*, 11.
- [12] Gautam, C. S., Pandey, P. (2023). An improving query optimization process in Hadoop MapReduce using ACO-genetic algorithm and HDFS MapReduce technique. *International Journal of Current Engineering and Technology*, 13, doi: 10.14741/ijcet/v.13.2.8.
- [13] Gautam, C. S., Pandey, P. (2022). A review on genetic algorithm models for Hadoop MapReduce in big data. *International Journal of Recent Scientific Research*, 13, 771–775. doi: 10.24327/ijrsr.2022.1303.0166.
- [14] Gautam, C. S., Pandey, P. (2019). A review of big data environment, tools and challenges. *Journal of Emerging Technologies and Innovative Research*, 6, 569–575, doi: 10.1729/Journal.22507.
- [15] Gautam, C. S., Wao, A. A. (2018). Speedup query processing in Hadoop using MapReduce framework. *Advances in Technology and Science*, 2, 43–53, doi: 10.31058/j.data.2018.21004.
- [16] Chaudhari, S., Gautam, C. S., Wao, A. A. (2024). Optimizing heart disease prediction accuracy using machine learning models. *International Journal of All Research Education and Scientific Methods*, 6, 1091–1098.
- [17] Chaudhari, S., Gautam, C. S., Wao, A. A. (2024). Enhancing heart disease prediction accuracy: A comparative study of machine learning models with ensemble method. *International Journal of Research and Innovation in Engineering (IJRIIE)*, 10, 4827–4833.