DOI: http://dx.doi.org/10.26483/ijarcs.v16i2.7234

Volume 16, No. 2, March-April 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

# AI-DRIVEN CLOUD INFRASTRUCTURE: ADVANCES IN KUBERNETES AND SERVERLESS COMPUTING

Vivek Sharma
System Analyst,MITS Deemed University
Gwalior, Madhya Pradesh ,India

***Abstract*—**Artificial Intelligence has been integrated into cloud infrastructure, making it revolutionizing modern computing by automating, scaling, and efficiency. The first of these is Kubernetes and the second is serverless computing. Kubernetes, a container orchestration platform, benefits from AI-driven enhancements in workload scheduling, auto-scaling, and resource optimization. By combining AI based predictive analytics with container deployment, overhead is reduced in terms of the operational overhead as well as the fault tolerance. However, serverless computing takes away the management of infrastructure leaving the developers free to write application logic. Serverless architectures with AI-based power can scale and adaptively allocate the resources and execute cloud workloads with minimum cost. This review study examines the latest development of AI-based Kubernetes and serverless computing, their influence on the cloud infrastructure. AI is discussed insofar as its orchestration role can be optimized, security is improved and self-healing cloud environments are made possible. The paper also studies challenges: latency, security issues, and uncertainties of integration of AI models. Through the analysis of state-of-the-art innovation and future trends, this review covers how AI is impacting the future cloud native computing.

***Keywords*—**AI-Driven, Cloud Infrastructure, Kubernetes, Serverless Computing, Cloud-Native, Orchestration

## I. INTRODUCTION

Cloud computing is a form of provision of on-demand computing resources to users of any kind over the internet that provides the users with the ability to use and manage resources in the cloud. It has changed how organizations that deploy, scale and manage their IT infrastructure do business [1]. In this regard, cloud services offer businesses extreme flexibility, scalability and cost efficiency; businesses can instead concentrate on their core competencies and the cloud service providers handle the complexities of IT infrastructure management.

The businesses are driven to move forward by organizations that want their businesses to advance, and this is happening in a distributed way, using a distributed computing architecture [2]. This shift is also changing the role of a Chief Technology Officer (CTO) and a Chief Information Officer (CIO). Although many still talk about the trend of digital transformation, and AI, ML, and extended reality (XR) are popular, cloud transformation is relevant, even if it is not the focus of mainstream media. Cloud computing continues to be a hot discussion point behind the scenes, and organizations are sensibly understanding that cloud adoption is not an option but a necessity.

The rise of Kubernetes is a major leap forward in cloud computing as a legal technology for serverless platforms. Open source serverless computing frameworks are well supported by various Kubernetes and are used to manage containerized applications. Kubernetes has been proven to be able to maintain performance levels similar to bare metal in old high-performance computing (HPC) workloads [3].

It is a relatively new paradigm in which developers can build and deploy their applications without having to set up underling infrastructures in traditional cloud river spaces [4][5]. Serverless computing reduces the efforts of programmers in maintaining the system by automating almost all system administration tasks. Initially, Amazon EC2 attracted developers by having full control of application instances to the point that developers had to manage and scale application instances heavily. Serverless computing solves this issue by removing the infrastructure complexities and using it to deploy the application more efficiently.

Edge computing is another emerging trend that has been growing steadily in the last few years. The paradigm of edge computing is an extension of cloud infrastructure where the computing resources are placed closer to the end user. They are general-purpose servers to specific types of IoT devices such as cameras and sensors that can collect and process data [6][7][8]. This is due to the fact that it reduces latency and processes the data near the production nodes as well as ensuring improved data security. Optimizing serverless computing parameters is essential for achieving high performance in edge environments.

As AI-driven cloud infrastructure advances, Kubernetes and the development of cloud-native apps are significantly influenced by serverless computing [9]. Understanding their impact on performance, scalability, and resource optimization is crucial for organizations looking to leverage these technologies effectively.

### *A. Structured of the paper*

The structure of this paper is as follows: Section II Role of AI in Cloud Infrastructure. Section III Kubernetes and Serverless: AI-Driven Orchestration. Section IV reviews literature and case studies, and Section V concludes with future directions.

## II. ROLE OF AI IN CLOUD INFRASTRUCTURE

Cloud-based infrastructures are seen to be the most appropriate for meeting resource and service demands [10]. However, there are several obstacles to transferring the enormous amounts of data produced to the cloud, including excessive latency, network congestion, and privacy concerns. Thankfully, edge computing has become a viable concept that places computing closer to the data source, reducing latency, burdening the cloud, and improving privacy, smart application quality of service, and user experience.

Integrating AI and cloud/edge computing fully unleashes their potential values, leading to a new intelligence computing paradigm. However, there are still many open-ended challenges during its implementation, such as limited computing, networks, and energy resources, accompanied by serious security issues. Meanwhile, the dynamic features of cloud/edge environments also complicate matters [11][12]. By learning from data and making decisions grounded in discernment, among other methods, such techniques can lead to machines with the capabilities to solve complicated problems. Therefore, advanced computational intelligence exhibits great promise and abundant prospects for applications in cloud/edge computing, which can serve as both an enabler that bolsters the service capabilities and as a problem-solver, surmounting the obstacles during the system design. This Special Issue endeavors to assemble scholarly studies that explore the paths of synergizing cloud/edge computing with advanced computational intelligence to guide the development of next-generation network technology [13].

There are many different uses for cloud computing's adaptability. It enables remote patient monitoring and telemedicine in the healthcare industry, where real-time data informs crucial medical choices. By enabling traffic optimization systems, it lowers traffic and enhances urban mobility in smart cities. Cloud systems in industrial environments serve as centers for predictive maintenance, increasing productivity and reducing downtime [14][15]. Still, there are difficulties. There are persistent challenges due to high bandwidth prices, latency problems, and the intricacies of handling data sovereignty regulations. To overcome these obstacles, solutions including hybrid architectures that integrate cloud and edge computing and data centers powered by renewable energy sources are being developed. As IoT and AI ecosystems develop, cloud computing is a key factor that is opening the door for intelligent systems that are effective, safe, and long-lasting. Its full potential may be achieved by utilizing its advantages and removing current obstacles.

## III. Benefits Of AI-Driven Cloud Solutions

The following are some advantages of AI-driven cloud solutions:

- **Enhanced Resource Management and Cost Efficiency:** The digital business environment has been completely transformed by cloud computing, which offers unmatched scalability, flexibility, and efficiency in service delivery and data management [16]. A number of important obstacles remain in spite of these developments, including maintaining high performance, optimizing cloud resources, and attaining cost-effectiveness. AI, a key component of cloud computing capabilities, has been recognised by recent technical breakthroughs as offering creative solutions for more efficient resource management.
- **Improved Security Measures:** In many industries (such as healthcare, agriculture, education, transportation, etc.), the use of new, emerging technologies like the IoT, wireless sensor networks (WSNs), cloud/edge computing, and 5G/6G communication networks can lead to numerous opportunities to improve people's quality of life and create intelligent systems that provide customers with innovative, high-quality services [17].
- **Accelerated Software Development Processes:** As technology has advanced and software needs have become more complicated, software engineering processes have undergone a fast and revolutionary change. Methodologies for software development have continuously improved since their inception, moving from the conventional waterfall model to more agile and iterative approaches.
- **Support for Industrial Internet of Things Applications:** IIoT has become a key component of the continuous digital transformation of companies, allowing for improved automation, intelligence, and connection in operations. In the industrial, healthcare, energy, and logistics industries, IIoT enables real-time monitoring, enhanced productivity, and creative business models by connecting physical systems with cutting-edge digital technology. Cloud computing is becoming a crucial facilitator of these systems as managing and analyzing the enormous amounts of data produced by linked devices is crucial to the IIoT's success [18][19].

## IV. Kubernetes And Serverless Computing: AI-Driven Orchestration

A wide range of interrelated issues influence manufacturing's capacity to satisfy the needs of today's dynamic markets in its constantly changing landscape [20][21]. These difficulties include a wide range of topics, all of which are essential to achieving operational excellence and long-term growth. Product quality, process quality, process planning and scheduling, adaptability in the context of market dynamics, human-machine interaction, and data quality and accessibility could all be included in a thorough analysis of the aforementioned challenges.

An open-source technology called Kubernetes makes it easier to automate containerized application deployment, scalability, and administration. It frees developers from worrying about the underlying infrastructure so they can concentrate on creating and implementing their applications. Users define desired application states using Kubernetes' declarative application management technique, and the system keeps track of them. Additionally, it offers strong application management and monitoring features, such as self-healing techniques for automated failure detection and recovery. All things considered, Kubernetes provides a strong and adaptable way to manage containerized apps in production settings [22].

### *A. Kubernetes and Its Role in Cloud Computing*

An open-source framework called Kubernetes, or K8s, was created to automate the deployment, scaling, and administration of containerized applications. It offers a platform that is both portable and expandable, making it possible to deploy and maintain applications across a cluster of computers in an effective manner. A pod serves as the basic unit of deployment in the context of container-based virtualization, especially in the context of the Kubernetes ecosystem. One or more closely related containers that are collocated and share resources and network namespaces make up a pod, which operates as a logical grouping [23].

### *B. AI-Enhanced Workload Orchestration*

The cloud computing paradigm was formed and developed via the introduction of pay-per-use and on-demand resource allocation methods brought about by the widespread use of virtualization. The difficulty of planning and implementing an effective coordinated execution of many virtual machines to optimize expenses for the owner has arisen due to the economic implications of resource utilisation. To use the elasticity offered by cloud computing, a multitude of orchestration solutions are

available for centrally managing applications with specific auto-scaling and QoS needs [24]. These orchestration platforms struggle to satisfy the stringent QoS standards set out by the service owners, notwithstanding their limited understanding of the underlying workings of the behavior and characteristics of the service.

This paradigm's relatively homogeneous supporting hardware, which is typically found in data centers, is one of its characteristics. In most cases, every machine in a cloud deployment has the same settings. For example, all of the machines have the same ad hoc operating systems, which may have been modified by vendors to improve global performance based on the kind of service offered. Cloud computing also frequently guarantees high network bandwidth and dependability. However, there has been a noticeable shift in recent years towards edge computing, which favors computation in more localized settings [25].

### *C. Emerging Trends in Kubernetes*

Kubernetes develops further through multiple new trends that improve its features. It can enhance Kubernetes through Operators that automate stateful application handling and turn operational human knowledge into functional software.

The following are the Emerging Trends in Kubernetes:

- **Kubernetes Operators:** Automates stateful application management by embedding operational knowledge into software [26].
- **Multi-Cloud Kubernetes:** Enables deployment across several cloud providers to improve dependability and prevent vendor lock-in [27].
- **AI-Driven Orchestration:** Uses machine learning for intelligent workload scheduling, auto-scaling, and anomaly detection.
- **Edge Computing Integration:** Expands Kubernetes to edge environments for low-latency processing and distributed computing.
- **Serverless Kubernetes:** Supports event-driven, auto-scaling applications without manual infrastructure management [28].

### *D. Serverless Computing*

A new and appealing paradigm for cloud application deployment is serverless computing, or simply serverless. This is mostly because business application architectures have recently shifted to microservices and containers. Serverless computing gives cloud providers more control over the whole development stack, lowers operating costs through effective resource optimization and management, offers a platform that promotes the use of other services in their ecosystem, and lessens the effort needed to create and maintain cloud-scale applications [29][30].

### *E. Traditional vs. Serverless Computing Technology*

Cloud computing has revolutionized how organizations approach infrastructure, application deployment, and service delivery. The following are the various Traditional vs. Serverless computing technologies shown in Figure 1:

Figure 1 contrasts traditional and serverless architectures. In a traditional setup, a client-side application typically interacts with a monolithic backend encompassing front-end logic, back-end logic, security measures, and a database. Conversely, serverless architecture, leveraging client-side logic and third-party services, distributes these components. The client application directly interacts with specific, independent services for security, back-end logic (often as Functions as a Service - FaaS), and databases, potentially reducing the operational overhead associated with managing dedicated servers.

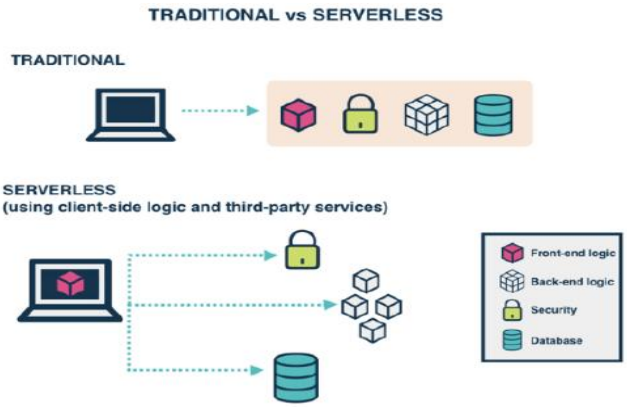


Fig. 1. Traditional vs. Serverless Computing Technology

- **Virtual Management:** Traditional cloud computing typically involves managing virtualized servers or containers where applications run, requiring a hands-on approach to provisioning, scaling, and maintenance. In contrast, serverless computing abstracts away the underlying infrastructure, freeing developers from worrying about servers so they can concentrate on code and functionality [31][32].
- **Infrastructure Management:** Traditional models necessitate manual provisioning and scaling, whereas serverless computing abstracts these responsibilities to the cloud provider.
- **Scalability:** Serverless architectures inherently support automatic scaling in response to incoming events, unlike traditional models that may require manual intervention.
- **Cost Structure:** Traditional models often involve paying for pre-allocated resources, potentially leading to underutilization. Serverless computing charges depending on real consumption, providing a more economical solution for workloads that fluctuate[33].

### *F. Key Benefits of Serverless Cloud Computing*

There are several benefits of serverless computing Some are as follows:

- **Improved Development Speed:** Software development is accelerated by serverless architectures. Development cycles are significantly shortened when developers can concentrate only on producing and improving code rather than managing infrastructure [34].
- **Scalability:** Scalability is a feature of serverless computing by nature. Without human involvement, cloud providers ensure smooth operation by automatically handling scalability in response to application demand.
- **Agility:** Independent development and deployment of microservices enable faster release cycles and more responsive adaptation to changing business requirements.
- **Portability:** Containerization ensures that applications run consistently across various environments, reducing deployment issues and enhancing portability.
- **Cost Efficiency:** The cost-effectiveness of serverless computing is among its most alluring features[35].

*G. Challenges of Serverless Computing*

Serverless computing faces a number of difficulties. A few of them are:

- **Software engineering challenges:** Some of the most important problems with the serverless paradigm have been found to be software engineering obstacles, such developer experience.
- **Data Management:** Ensuring data consistency and integrity across distributed services can be challenging, necessitating careful design and coordination.
- **Security:** The increased surface area due to numerous services and APIs can elevate security risks, demanding robust authentication, authorization, and monitoring mechanisms.
- **Resource Overhead:** While containers are lightweight, the overhead of running numerous instances can accumulate, impacting resource utilization and costs.
- **System (operational) challenges:** Cloud functions are extremely dynamic, which creates system challenges that necessitate improvements in cloud function lifecycle management, cost predictability, and security [36].

*H. Kubernetes for Serverless Computing*

Serverless computing is made possible and improved by Kubernetes, which offers scalability, flexible, and efficient container orchestration. Below are key aspects of how Kubernetes contributes to serverless computing:

- **Serverless Frameworks –** Supports Knative, OpenFaaS, and Kubeless for deploying serverless applications.
- **Autoscaling –** Uses Horizontal Pod Autoscaler (HPA) and KEDA for dynamic function scaling.
- **Cost Efficiency –** Enables a pay-per-use model by allocating resources only when needed.
- **Multi-Cloud & Portability –** Avoids vendor lock-in and supports hybrid & multi-cloud deployments.
- **Event-Driven Execution –** Integrates with Kafka, NATS, and cloud events for real-time processing.
- **Security & Isolation –** Implements RBAC, network policies, and pod security for secure execution.[37]

*I. Kubernetes VS Serverless*

These two tools work differently to simplify deployment tasks, although they exist for separate application purposes.

- Organizations choose Kubernetes to manage applications that depend on multistage deployment while maintaining stateful services or managing their infrastructure.
- Serverless technology fits with applications that run on events and microservices plus services that need to grow quickly or need low maintenance efforts[38].

The complexity of the project and the experience of crew will determine whether you select Kubernetes or serverless [39].

## V. LITERATURE REVIEW

In this section offers a thorough analysis of the research on Kubernetes and Serverless Computing Developments with AI-Powered Cloud Infrastructure, with a summarized overview presented in Table I.

Ma et al. (2025) a neural-enhanced interference-aware resource provisioning system for serverless computing. They model the resource provisioning of serverless functions as a novel combinatorial optimization problem, wherein the constraints on the queries per second are derived from the neural network performance model. By leveraging neural networks to model the nonlinear performance fluctuations under various interference sources, their approach better captures the real-world behavior of serverless functions [40].

Ciptaningtyas et al. (2024) Using the Knative Framework, users can implement serverless services without relying on other cloud providers. This research aims to implement a Serverless system using Knative and conduct performance testing with parameters such as failure rates, response times, and resource use. The study produces a viable monitoring system and a working Serverless system using Kubernetes. It is discovered that serverless CPU and RAM function similarly to traditional systems in terms of resource utilisation [41].

Rahman et al. (2024) proposes a novel artificial intelligence (AI)–driven IT infrastructure backed by blockchain technology, specifically designed to optimize risk management processes in diverse organizational environments. By leveraging artificial intelligence for predictive analytics, anomaly detection, and data-driven decision-making, combined with blockchain's secure and immutable ledger for data integrity and transparency, the proposed infrastructure offers a robust solution to existing challenges in risk management. The infrastructure is adaptable and scalable to support a variety of risk management methodologies, providing a more secure, efficient, and intelligent system. The findings highlight significant improvements in the accuracy, speed, and reliability of risk management, underscoring the infrastructure's capability to proactively address emerging cyber threats [42].

Pranata, Wijayanto and Fajar Sidiq (2023) This document explores the use of AI-driven optimization tools in cloud and edge environments with respect to their potential application in the dynamic allocation of resources, the management of workloads and the lowering of latency. AI can intelligently organize activities by utilizing ML models and predictive analytics, allowing systems to grow with equal performance. Based on the approval of dynamic workload distribution, energy efficient resource allocation, as well as real time decision making in adaptive scaling, this study analyzes the key optimization strategies [43].

Tuli et al. (2023) investigate the possible applications of AI-driven optimization tools in cloud and edge contexts for workload management, latency reduction, and dynamic resource allocation. AI can intelligently plan tasks using ML models and predictive analytics, enabling systems to grow with comparable performance. This paper examines the main optimization techniques based on the acceptance of dynamic workload distribution, energy-efficient resource allocation, and real-time decision-making in adaptive scaling [44].

Dehury and Srirama (2024) explore the challenges associated with this integration, especially with Apache Spark. Despite their advanced capabilities, modern SDPEs still lack full maturity in terms of efficiently managing dynamic resource demands and seamlessly integrating with other technologies. To fill this gap, it proposes the architecture of ISIM-SDP, acronym for Integrating Serverless and DRL for Infrastructure Management in Streaming Data Processing across edge-cloud continuum. By implementing a DRL-based approach, the system dynamically adjusts resource allocation in real time, enhancing the flexibility and scalability of computational resources [45].

TABLE I. SUMMARY OF KUBERNETES AND SERVERLESS COMPUTING WITH AI-DRIVEN CLOUD INFRASTRUCTURE

| Reference | Focus On | Key Findings | Challenges | Limitations/Future Gap |
|---|---|---|---|---|
| Ma et al. (2025) | Neural-enhanced, interference-aware resource provisioning in serverless computing | Models serverless function provisioning as a combinatorial optimization problem using neural networks to predict performance under interference | Accurately modeling interference in dynamic environments | Need for further validation in diverse real-world serverless setups |
| Ciptaningtyas et al. (2024) | Serverless deployment using Knative and Kubernetes | Developed a functioning serverless system with monitoring capabilities; resource use (CPU, RAM) is comparable to traditional systems | Ensuring reliability and performance in self-managed serverless setups | Lacks detailed comparative analysis with commercial serverless platforms |
| Rahman et al. (2024) | AI-driven IT infrastructure with blockchain for risk management | Improved accuracy, speed, and reliability in handling cyber threats; scalable and adaptive system | Balancing performance and security in hybrid infrastructures | Implementation complexity and integration in legacy systems |
| Pranata, Wijayanto and Sidiq (2023) | AI-based optimization in cloud/edge resource allocation | AI enables dynamic resource allocation, latency reduction, and energy efficiency | Coordinating AI models across heterogeneous environments | Requires extensive data and continuous model retraining |
| Tuli et al. (2023) | AI optimization in workload management and scaling | AI supports real-time decision-making for adaptive scaling and resource optimization | Managing unpredictable workload patterns in distributed systems | Limited evaluation in edge-heavy deployments |
| Dehury and Srirama (2024) | DRL-based resource management in SDPE with Apache Spark | ISIM-SDP enables real-time, scalable resource allocation using DRL | Integration with existing stream processing ecosystems | Full integration with modern SDPEs and cross-platform generalization is yet to be achieved |

## VI. CONCLUSION AND FUTURE WORK

The integration of AI into cloud infrastructure, particularly within Kubernetes and serverless computing, is reshaping the deployment, management, and optimization of applications. AI-driven orchestration enhances workload efficiency, automates resource scaling, and improves fault tolerance in Kubernetes environments. Similarly, in serverless computing, AI enables intelligent function scaling, cost optimization, and adaptive workload management, leading to more efficient and resilient cloud services. These advancements reduce operational complexity while improving performance, security, and cost-effectiveness in cloud-native applications. Despite these benefits, challenges persist. AI integration into cloud orchestration demands significant computational resources, raising concerns about latency and cost efficiency. Security risks, including AI model vulnerabilities and data privacy issues, require continuous attention. Furthermore, interoperability across diverse cloud environments remains a critical area for improvement.

Future research should focus on enhancing AI models for real-time decision-making in cloud environments. Investigating AI-powered self-healing mechanisms, adaptive workload prediction models, and energy-efficient cloud management strategies will be crucial. Additionally, advancements in federated learning and privacy-preserving AI can address security concerns in cloud-native applications. As AI develops further, its combination with serverless computing and Kubernetes will open the door to more self-sufficient, effective, and intelligent cloud infrastructures.

## REFERENCES


[1] K. Anbalagan, "AI in Cloud Computing: Enhancing Services and Performance," Int. J. Comput. Eng. Technol., vol. 15, pp. 622–635, 2024, doi: 10.5281/zenodo.13353681.

[2] K. Govindan et al., "Industry Surveys IT Consulting & Other Services.," J. Clean. Prod., 2018.

[3] H. Chahed et al., "AIDA—A holistic AI-driven networking and processing framework for industrial IoT applications," Internet of Things (Netherlands), 2023, doi: 10.1016/j.iot.2023.100805.

[4] J. Decker, P. Kasprzak, and J. M. Kunkel, "Performance Evaluation of Open-Source Serverless Platforms for Kubernetes," Algorithms, 2022, doi: 10.3390/a15070234.

[5] B. K. R. Janumpally, "A Review on Data Security and Privacy in Serverless Computing: Key Strategies, Emerging Challenges," Int. J. Innov. Sci. Res. Technol., vol. 10, no. 3, p. 9, 2025.

[6] H. Martins, F. Araujo, and P. R. da Cunha, "Benchmarking Serverless Computing Platforms," J. Grid Comput., 2020, doi: 10.1007/s10723-020-09523-1.

[7] A. Goyal, "Optimising Cloud-Based CI/CD Pipelines: Techniques for Rapid Software Deployment," Tech. Int. J. Eng. Res., vol. 11, no. 11, pp. 896–904, 2024.

[8] J. Thomas, "The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains," Int. J. Res. Anal. Rev., vol. 8, no. 3, pp. 874–878, 2021.

[9] T. Rausch, A. Rashed, and S. Dustdar, "Optimized container scheduling for data-intensive serverless edge computing," Futur. Gener. Comput. Syst., 2021, doi: 10.1016/j.future.2020.07.017.

[10] V. Prajapati, "Cloud-Based Database Management: Architecture, Security, challenges and solutions," J. Glob. Res. Electron. Commun., vol. 01, no. 1, pp. 07–13, 2025, doi: https://doi.org/10.5281/zenodo.14934833.

[11] N. Malali, "Adversarial Robustness of AI-Driven Claims Management Systems," Int. J. Adv. Res. Sci. Commun. Technol., 2025.

[12] S. Arora, S. R. Thota, and S. Gupta, "Artificial Intelligence-Driven Big Data Analytics for Business Intelligence in SaaS Products," in 2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), IEEE, Aug. 2024, pp. 164–169. doi: 10.1109/IC2SDT62152.2024.10696409.

[13] S. U. Amin and M. S. Hossain, "Edge Intelligence and Internet of Things in Healthcare: A Survey," IEEE Access, 2021, doi: 10.1109/ACCESS.2020.3045115.

[14] V. Gharibvand et al., "Cloud based manufacturing: A review of recent developments in architectures, technologies, infrastructures, platforms and associated challenges," International Journal of Advanced Manufacturing Technology. 2024. doi: 10.1007/s00170-024-12989-y.

[15] P. M. Rajendra Prasad Sola, Nihar Malali, "Cloud Database Security: Integrating Deep Learning and Machine Learning


 

for Threat Detection and Prevention: 0," Notion Press, 2025.

[16] S. Murri, "Data Security Environments Challenges and Solutions in Big Data," Int. J. Curr. Eng. Technol., vol. 12, no. 6, pp. 565–574, 2022.

[17] J. Kumar Chaudhary, S. Tyagi, H. Prapan Sharma, S. Vaseem Akram, D. R. Sisodia, and D. Kapila, "Machine Learning Model-Based Financial Market Sentiment Prediction and Application," in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, May 2023, pp. 1456–1459. doi: 10.1109/ICACITE57410.2023.10183344.

[18] S. A. Ionescu and V. Diaconita, "Transforming Financial Decision-Making: The Interplay of AI, Cloud Computing and Advanced Data Management Technologies," Int. J. Comput. Commun. Control, 2023, doi: 10.15837/ijccc.2023.6.5735.

[19] S. Arora and S. R. Thota, "Ethical Considerations and Privacy in AI-Driven Big Data Analytics," Int. Res. J. Eng. Technol., vol. 11, no. 05, 2024.

[20] A. Gogineni, "Multi-Cloud Deployment with Kubernetes: Challenges, Strategies, and Performance Optimization," Int. Sci. J. Eng. Manag., vol. 1, no. 02, 2022.

[21] J. Thomas, "Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics," J. Emerg. Technol. Innov. Res., vol. 8, no. 9, pp. 357–364, 2021.

[22] T. Subramanya and R. Riggio, "Centralized and Federated Learning for Predictive VNF Autoscaling in Multi-Domain 5G Networks and beyond," IEEE Trans. Netw. Serv. Manag., 2021, doi: 10.1109/TNSM.2021.3050955.

[23] C. Perducat, D. C. Mazur, W. Mukai, S. N. Sandler, M. J. Anthony, and J. A. Mills, "Evolution and Trends of Cloud on Industrial OT Networks," IEEE Open J. Ind. Appl., 2023, doi: 10.1109/OJIA.2023.3309669.

[24] V. Pillai, "Integrating AI-Driven Techniques in Big Data Analytics: Enhancing Decision-Making in Financial Markets," Int. J. Eng. Comput. Sci., vol. 12, no. 7, 2023.

[25] E. Lee et al., "Device Description in HYDRA Middleware," IEEE Access, 2021.

[26] A. Gogineni, "Chaos Engineering in the Cloud-Native Era: Evaluating Distributed AI Model Resilience on Kubernetes," J Artif Intell Mach Learn Data Sci 2024, vol. 3, no. 1, pp. 2182–2187, 2025.

[27] S. Padmakala, M. Al-Farouni, D. D. Rao, K. Saritha, and R. P. Puneeth, "Dynamic and Energy-Efficient Resource Allocation using Bat Optimization in 5G Cloud Radio Access Networks," in 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), IEEE, Aug. 2024, pp. 1–4. doi: 10.1109/NMITCON62075.2024.10699133.

[28] C. Carrión, "Kubernetes as a Standard Container Orchestrator - A Bibliometric Analysis," J. Grid Comput., 2022, doi: 10.1007/s10723-022-09629-8.

[29] H. Shafiei, A. Khonsari, and P. Mousavi, "Serverless Computing: A Survey of Opportunities, Challenges, and Applications," ACM Comput. Surv., 2022, doi: 10.1145/3510611.

[30] S. S. S. Neeli, "Cloud Migration DBA Strategies for Mission-Critical Business Applications," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 11, pp. 591–598, 2023.

[31] V. Prajapati, "Role of Identity and Access Management in Zero Trust Architecture for Cloud Security : Challenges and Solutions," pp. 6–18, 2025, doi: 10.48175/IJARSCT-23902.

[32] V. Pillai, "Anomaly Detection in Financial and Insurance Data-Systems," J. AI-Assisted Sci. Discov., vol. 4, no. 2, 2024.

[33] J. M. O. Candel, A. Elouali, F. J. M. Gimeno, and H. Mora, "Cloud vs Serverless Computing: A Security Point of View," in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-3-031-21333-5_109.

[34] J. L. Deepak Dasaratha Rao, Sairam Madasu, Srinivasa Rao Gunturu, Ceres D'britto, "Cybersecurity Threat Detection Using Machine Learning in Cloud-Based Environments: A Comprehensive Study," Int. J. Recent Innov. Trends Comput. Commun., vol. 12, no. 1, 2024.

[35] D. Loconte, S. Ieva, A. Pinto, G. Loseto, F. Scioscia, and M. Ruta, "Expanding the cloud-to-edge continuum to the IoT in serverless federated learning," Futur. Gener. Comput. Syst., 2024, doi: 10.1016/j.future.2024.02.024.

[36] H. B. Hassan, S. A. Barakat, and Q. I. Sarhan, "Survey on serverless computing," Journal of Cloud Computing. 2021. doi: 10.1186/s13677-021-00253-7.

[37] S. K. Mondal, R. Pan, H. M. D. Kabir, T. Tian, and H. N. Dai, "Kubernetes in IT administration and serverless computing: An empirical study and research challenges," J. Supercomput., 2022, doi: 10.1007/s11227-021-03982-3.

[38] M. S. Samarth Shah, "Deep Reinforcement Learning For Scalable Task Scheduling In Serverless Computing," Int. Res. J. Mod. Eng. Technol. Sci., vol. 3, no. 12, pp. 1845–1852, 2021, doi: DOI : https://www.doi.org/10.56726/IRJMETS17782.

[39] M. Aazam, S. Zeadally, and K. A. Harras, "Fog Computing Architecture, Evaluation, and Future Research Directions," IEEE Commun. Mag., vol. 56, no. 5, pp. 46–52, May 2018, doi: 10.1109/MCOM.2018.1700707.

[40] R. Ma, Y. Zhan, C. Wu, Z. Hong, Y. Ali, and Y. Xia, "Qora: Neural-Enhanced Interference-Aware Resource Provisioning for Serverless Computing," IEEE Trans. Autom. Sci. Eng., pp. 1–16, 2025, doi: 10.1109/TASE.2025.3526197.

[41] H. T. Ciptaningtyas, R. R. Hariadi, F. D. Rosyadi, and S. S. Al Azmi, "Serverless Computing Model Using Kubernetes and Knative in a Scalable Cloud Development," in 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C), 2024, pp. 659–664. doi: 10.1109/BTS-I2C63534.2024.10942132.

[42] M. M. Rahman, B. P. Pokharel, S. A. Sayeed, S. K. Bhowmik, N. Kshetri, and N. Eashrak, "riskAIchain: AI-Driven IT Infrastructure—Blockchain-Backed Approach for Enhanced Risk Management," Risks, vol. 12, no. 12, p. 206, Dec. 2024, doi: 10.3390/risks12120206.

[43] M. Pranata, A. Wijayanto, and M. Fajar Sidiq, "Serverless Autoscaling Metrics for Optimum Performance on Edge Computing," in 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2023, pp. 65–69. doi: 10.1109/ISRITI60336.2023.10467288.

[44] S. Tuli et al., "AI augmented Edge and Fog computing: Trends and challenges," Journal of Network and Computer Applications. 2023. doi: 10.1016/j.jnca.2023.103648.

[45] C. K. Dehury and S. N. Srirama, "Integrating Serverless and DRL for Infrastructure Management in Streaming Data Processing across Edge-Cloud Continuum," in 2024 IEEE 44th International Conference on Distributed Computing Systems Workshops (ICDCSW), 2024, pp. 93–101. doi: 10.1109/ICDCSW63686.2024.00020.

 