DISTRIBUTED SYSTEMS ANALYSIS FOR ARTIFICIAL INTELLIGENCE IN CLOUD COMPUTING: A COMPREHENSIVE REVIEW OF AI-BASED APPLICATIONS AND SERVICES

Muhammad Ubaidurrohman[1]

[1] *Universitas Dian Nuswantoro (UDINUS), Indonesia*

GENERATE

Breakthroughs Information Technology

e-ISSN: 3109-8495
Vol 02 (1) 2026 p. 1-19

Coresponding author:
Muhammad Ubaidurrohman
Email :
muhammad.ubaidurrohman@gmail.com

Received 19 June 2026;
Accepted 9 July 2026;
Published 15 July 2026.

Conflict of interest statement: Author(s) reported no conflict of interest

DOI: http://doi.org/10.70764/gdpu-bit.2026.2(1)-01

ABSTRACT

Objective: This study aims to examine the development and application of Artificial Intelligence (AI) in distributed cloud environments and identify emerging technologies, research trends, benefits, challenges, and future directions of AI-driven cloud services.

Research Design & Methods: A literature review was conducted using the Scopus database. The search focused on publications related to Artificial Intelligence, Distributed Cloud, Cloud Environment, and Applications and Services.

Findings: The results indicate a growing research interest in AI-enabled distributed cloud environments during the last five years. The literature reveals three major technological pillars: Optimization-Based AI, Machine Learning-Based AI, and Deep Learning-Based AI. AI has been widely adopted for resource management, workload scheduling, anomaly detection, cybersecurity, and autonomous cloud operations. Emerging trends include cloud-native architectures, multi-cloud systems, edge-cloud computing, Explainable AI (XAI), federated learning, and multi-agent reinforcement learning. However, challenges remain regarding infrastructure heterogeneity, scalability, security, explainability, and the generalization capability of AI models across diverse cloud environments.

Implications: The findings highlight the importance of integrating intelligent and adaptive AI mechanisms to improve efficiency, security, scalability, and operational automation in distributed cloud infrastructures.

Contribution & Value Added: This study provides a comprehensive synthesis of recent AI developments in distributed cloud environments and proposes a structured classification of AI approaches. The results offer insights into current research gaps and future opportunities for developing autonomous, trustworthy, and sustainable AI-driven cloud ecosystems.

Keywords: Artificial Intelligence, Distributed Cloud, Cloud Computing, Resource Management, Reinforcement Learning.

JEL codes: C45, C88, L86
Article type: Review paper

INTRODUCTION

Advances in digital technology over the past decade have led to massive data growth, increased computing demands, and the expanding use of artificial intelligence (AI)-based services (Zangana and Zeebaree, 2024). Organizations across sectors, such as healthcare, finance, manufacturing, education, and transportation, are increasingly relying on AI to support real-time decision-making, process automation, predictive analytics, and intelligent service delivery. This condition drives the need for computing infrastructure that is capable of providing resources flexibly, scalable, and efficiently.

Cloud computing has become a key foundation for supporting modern AI implementations, providing compute, storage, and networking resources that can be elastically accessed as needed. However, the increasing complexity of AI applications, particularly in the era of Generative AI and Large Language Models (LLMs), poses new challenges related to scalability, latency, resource coordination, data security, and processing efficiency in distributed cloud environments (Gill et al., 2022; Wu, 2020).

In this context, the integration between Artificial Intelligence and distributed systems emerges as a new computing paradigm that enables the utilization of geographically distributed computing resources while remaining effectively coordinated. The combination of these two technologies is an important catalyst in the transformation of cloud computing because it allows the system to adapt dynamically to changing workloads, increase data processing efficiency, and support the development of large-scale intelligent services (Duan et al., 2023).

Distributed cloud environments provide the ability to distribute computing loads across multiple nodes or data centers, thereby improving service availability, fault tolerance, and overall system performance. On the other hand, AI plays a role in optimizing resource management, predicting computing needs, automating decision-making, and improving the quality of cloud services. The synergy between distributed systems and AI has given birth to various innovations such as AI as a Service (AIaaS), intelligent cloud orchestration, autonomous resource management, predictive maintenance, smart healthcare systems, intelligent transportation systems, and cloud-edge intelligence for the Internet of Things (IoT) (Sandhu et al., 2022; Yang et al., 2020).

Previous research has shown that distributed computing plays an important role in supporting parallel processing and large-scale computing (Zebari and Yaseen, 2011). Abdullah et al. (2020) showed that integrating intelligent technology into an organization's resource management system can enhance information management and decision-making effectiveness. Furthermore, Abdulqader and Zeebaree (2021) found that distributed-memory parallel processing contributes significantly to improving the performance of computationally intensive applications running in cloud environments.

In addition, several recent studies have investigated the relationship between AI and distributed systems in various contexts. Duan et al. (2023) reviewed the development of Distributed Artificial Intelligence (DAI) as an approach to enhancing the capabilities of intelligent systems running in distributed environments. Wu (2020) discusses the integration of cloud-edge computing to support IoT services that require fast response and real-time data processing. Sandhu et al. (2022) highlight the importance of leveraging AI in enhancing cloud storage security through more adaptive threat detection and data protection mechanisms. Meanwhile, Gill et al. (2022) identified various future trends that indicate that AI will be a key component in the development of next-generation computing systems.

Although various studies have discussed AI, cloud computing, and distributed systems separately or partially, there are still limitations in comprehensive knowledge mapping regarding the development of AI applications and services in distributed cloud environments. Few studies systematically identify the architectures used, emerging application categories, available AI services, implementation challenges, and future research opportunities within a single, integrated framework. Therefore, a Systematic Literature Review (SLR) is needed that can synthesize current research findings to provide a comprehensive overview of the development of Artificial Intelligence in distributed cloud environments.

Based on these problems, this study aims to analyze and map the development of Artificial Intelligence research in distributed cloud environments by identifying publication trends, AI technologies used, characteristics of distributed cloud architectures that support AI implementation, categories of developed applications and services, benefits and challenges faced, and future research directions. Specifically, this study seeks to answer five research questions, namely: (RQ1) what are the trends in Artificial Intelligence research development in distributed cloud environments; (RQ2) what architectures and technologies are used in implementing AI in distributed cloud environments; (RQ3) what are the categories of AI-based applications and services that are developing; (RQ4) what are the benefits, challenges, and limitations of implementing AI in distributed cloud environments; and (RQ5) what are the research opportunities and future development directions in the field of Artificial Intelligence in distributed cloud environments. The results of this study are expected to provide a comprehensive understanding of the state of the art in AI research in distributed clouds and serve as a conceptual foundation for the development of next-generation AI services that are more adaptive, secure, scalable, and sustainable.

LITERATURE REVIEW

Foundations of Distributed Systems

A distributed system is a collection of independent computers connected through a network and working in a coordinated manner, appearing to users as a single, integrated system. These systems are designed to support large-scale data processing and management and possess key characteristics such as resource sharing, scalability, fault tolerance, concurrency, and transparency. In the development of modern computing, distributed systems are the foundation for cloud computing, edge computing, the Internet of Things (IoT), and Artificial Intelligence (AI). The integration of Distributed Artificial Intelligence (DAI) with cloud-edge infrastructure enables parallel and collaborative data processing across multiple computing nodes, improving processing efficiency, system scalability, and real-time decision-making capabilities (Duan et al., 2023).

Artificial Intelligence Basics

Artificial Intelligence (AI) is a branch of computer science that focuses on developing systems that can imitate human cognitive abilities, such as learning, reasoning, decision-making, and problem-solving through various approaches such as Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Computer Vision, and Reinforcement Learning (RL). AI is becoming a core technology in next-generation computing because it is able to process large amounts of data into valuable information to support automation, prediction, and intelligent decision-making (Gill et al., 2022). In cloud and distributed systems environments, AI is leveraged for resource optimization, workload prediction, anomaly detection, cybersecurity, and adaptive service automation. The development of AI has also given rise to new approaches such as Explainable AI (XAI), Federated Learning, and Generative AI, which play a crucial role in enhancing transparency, data privacy, and intelligent system capabilities in distributed computing ecosystems.

Distributed Cloud Computing, Edge Computing, and Artificial Intelligence

Distributed Cloud Computing is a development of cloud computing that distributes computing services and resources across multiple geographic locations to improve scalability, service availability, and reduce latency. Its development has given rise to various architectures such as multi-cloud, hybrid cloud, cloud-native, fog computing, and the cloud-edge continuum, all of which support more flexible workload management (Gill et al., 2022). On the other hand, edge computing moves some computing processes closer to the data source, thereby accelerating system response and supporting real-time IoT applications (Wu, 2020). Artificial Intelligence plays a role not only as a service, but also as an intelligent mechanism for predicting workloads, dynamic resource allocation, intelligent scheduling, and automatic orchestration, thus creating a more adaptive, efficient, and autonomous cloud system (Duan et al., 2023; Kunduru, 2023).

Security in Distributed AI Systems

Security is one of the main challenges in implementing AI in a distributed cloud environment because data and computing processes are spread across various nodes and geographical locations. Sandhu et al. (2022) showed that AI integration in cloud security systems can improve threat detection capabilities through behavioral analysis, anomaly detection, and security response automation. In addition, Xu et al. (2020) explained that the combination of AI and edge computing can improve the security of IoT services by detecting threats locally before data is sent to the cloud.

METHODS

This study uses a systematic literature review method designed to collect, critically evaluate, and synthesize scientific findings from previous research relevant to the topic of artificial intelligence in distributed systems in cloud computing environments. This methodological approach is rigorously structured to minimize bias and ensure the reliability of the analysis results.

Using a Systematic Literature Review (SLR) approach, this study conducts a comprehensive analysis of various studies discussing the integration of Artificial Intelligence (AI) in distributed cloud computing environments. The study focuses on the implementation of AI technologies, such as machine learning, deep learning, reinforcement learning, explainable AI, and intelligent optimization techniques, across various distributed computing architectures, including cloud-native, multi-cloud, edge-cloud, cloud-edge-IoT continuum, and hybrid cloud. Each study is analyzed based on the AI technology used, the characteristics of the distributed architecture, the application domain, the benefits gained, the implementation challenges, and the future development direction. The results of this literature synthesis are expected to provide a deeper understanding of the development trends of distributed cloud for AI services and serve as a basis for developing intelligent cloud architecture models in future research.

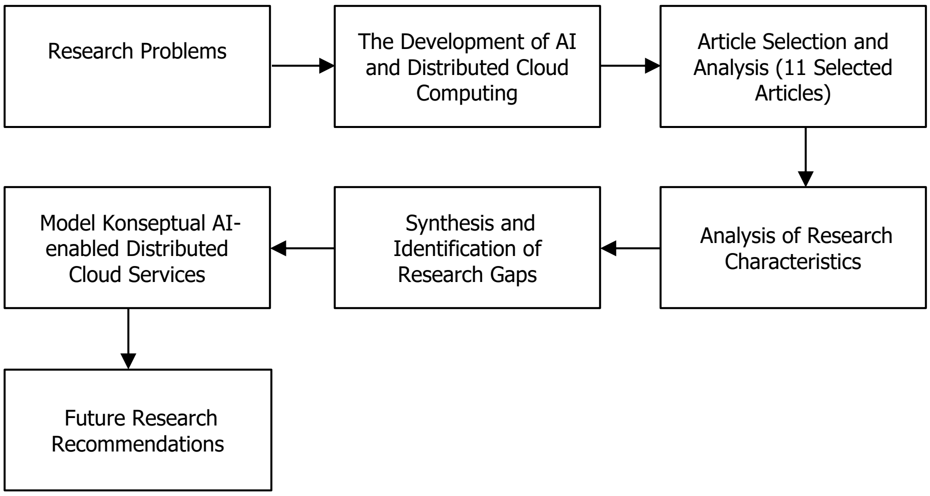

Figure 1. Research Flow

RESULT

A literature search was conducted in the Scopus database using keywords focused on Artificial Intelligence (AI) integration, Distributed Cloud, Cloud Environment, and Applications and Services. The search process was limited to English-language journal articles and conference papers that had reached the final publication stage. Based on the search results, 16 documents were obtained that were relevant to the research topic of the use of AI in distributed cloud computing environments, and only 11 studies were analyzed in figure 2.

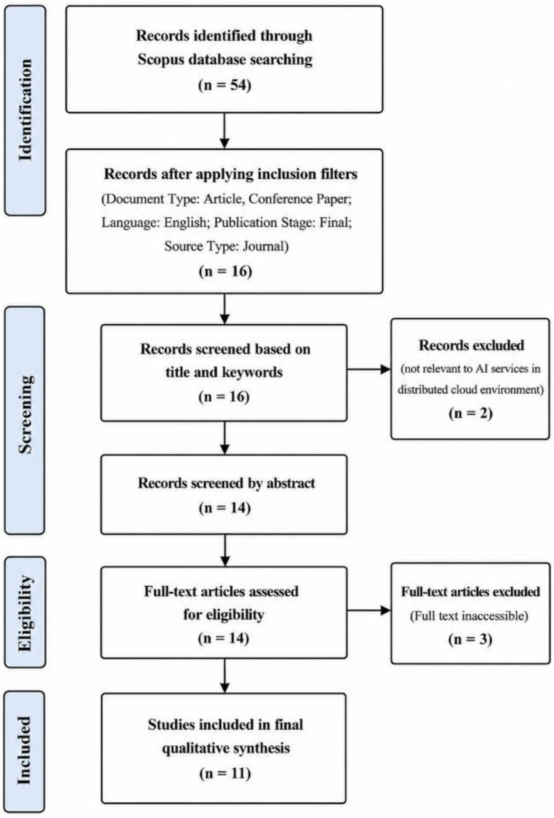

Figure 2. PRISMA Flow

Based on the distribution of publications per year, research on Artificial Intelligence in distributed cloud environments has shown an increasing trend over the past five years. Only one publication was found in 2021 and 2022, indicating that this topic is still in its early stages of development. The number of publications then increased to 4 documents in 2023, followed by 3 documents in 2024, 4 documents in 2025, and 3 documents in 2026. The increasing number of publications indicates the growing attention of academics and practitioners to the use of AI to support resource management, security, service optimization, and decision-making automation in distributed cloud environments. Furthermore, the development of supporting technologies such as edge computing, multi-cloud, cloud-native architecture, microservices, and the Internet of Things (IoT) has also driven the growing need for AI-based solutions that can work adaptively in increasingly complex computing environments.

In terms of country-of-origin publications, the United States was the largest contributor with five publications. This dominance demonstrates the country's strong research ecosystem and cloud technology industry, supported by global technology companies such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud, and VMware. India came in second with four publications, demonstrating significant growth in cloud computing and artificial intelligence research. Armenia also contributed three publications, most of which focused on cloud system observability, distributed tracing, and AI-based root cause analysis. Furthermore, research also came from other countries, including China, Spain, Bulgaria, Chile, Ghana, Indonesia, and others. This country distribution demonstrates that research on AI in the distributed cloud has become a global concern and is not limited to countries with high levels of technological development.

The search results show that the dominant research topics include cloud-native architecture, multi-cloud systems, edge-cloud computing, cloud-edge-IoT continuum, AI-based resource management, intrusion detection systems, Explainable AI (XAI), and reinforcement learning for cloud resource optimization. These findings suggest that AI integration in distributed cloud should focus not only on improving operational efficiency but also on the system's ability to automatically adapt to changing workloads, network conditions, and security threats. Overall, this field is growing rapidly with research increasingly moving towards AI-driven cloud management, federated learning, edge intelligence, explainable AI, the cloud-edge continuum, and sustainable cloud computing, thus potentially becoming the main foundation for next-generation digital services that are more intelligent, adaptive, secure, and scalable.

To obtain a comprehensive overview of research developments in the field of Artificial Intelligence Services in a Distributed Cloud environment, an in-depth analysis of articles obtained from the literature search was conducted. Based on the established inclusion and exclusion criteria, 11 key articles (state-of-the-art studies) were selected that were highly relevant to the research focus. Each article was analyzed based on the author, the AI technology used, the distributed cloud architecture implemented, the application domain, the benefits gained, and the challenges faced. The results of the classification and synthesis of the literature are presented in Table 1 as a basis for identifying research trends, research gaps, and opportunities for developing Artificial Intelligence Services in a Distributed Cloud environment.

Table 1. Classification of research articles

Author	AI Technology	Distributed Architecture	Application Domain	Benefits	Challenges
(Al-Tarawneh et al., 2025)	Multi-Criteria Decision Making (MCDM), TOPSIS, VIKOR, PROMETHEE, Hybrid Weighting Methods	Multi-provider Edge-Fog-Cloud Architecture	IoT, Internet of Vehicles (IoV), Industrial IoT (IIoT), Healthcare, Smart Cities, Home Automation, Environmental Monitoring	Improved task response time (67.6%), execution cost reduction (79.1%), lower energy consumption, enhanced security and fairness, scalable and robust task offloading, multi-objective optimization	Resource heterogeneity, dynamic workloads, QoS requirements, multi-provider resource allocation, latency, energy efficiency, security management
(Poghosyan et al., 2024)	Explainable AI (XAI), Rule-Induction Systems (RIPPER, C5.0), Machine Learning, Dempster-Shafer Theory (DST)	Native Cloud Applications, Microservices Architecture, Distributed Tracing (DT), Kubernetes/Docker-based Cloud Systems	Application Monitoring, AIOps, Root Cause Analysis (RCA), Cloud Application Troubleshooting, Performance Optimization	Enhanced observability, reduced troubleshooting time, improved trust and transparency, efficient trace management, reduced noise in diagnostics, faster execution time	Massive trace volumes, storage cost, loss of important traces during sampling, dynamic resource allocation, scalability issues, lack of AI transparency
(Xiao et al., 2022)	Game Theory (GT), Nash Equilibrium-based Scheduling, Peak Clipping Algorithm (PCA)	Distributed Cloud Environment with Cloud Provider–Customer Interaction Model	Cloud Computing, IoT Services, AI Service Scheduling, Resource Management	Reduced peak workload, improved QoS, balanced resource utilization, increased provider profit, reduced customer cost, fast convergence to Nash Equilibrium	Peak-load management, cloud provider overload, multi-party optimization, task scheduling complexity, large solution space, time-flexible resource allocation
(Talhar and Gaikwad, 2025)	Multi-Agent Reinforcement Learning (MARL), MADDPG, QMIX, Deep Predictive Modeling (LSTM), Evolutionary Optimization (Genetic Algorithm)	Decentralized Cloud Ecosystem, Multi-Agent Cloud Infrastructure, Autonomous VM-based Resource Management	Cloud Resource Provisioning, Cloud Resource Management, Sustainable Cloud Computing	Improved resource utilization (23%), reduced operational cost (21%), improved SLA compliance (>94%), reduced energy consumption (16.5%), increased throughput, adaptive and proactive resource management	Scalability limitations of traditional heuristics, DRL coordination challenges, centralized decision models, reactive resource management, multi-agent coordination complexity
(Appiahene et al., 2026)	Hybrid Deep Learning Architecture (GConvTrans), Graph Convolutional	Distributed Cloud Environment, Cloud-IoT Network Intrusion Detection System (NIDS)	Network Intrusion Detection, Cloud Security, IoT Security	High intrusion detection accuracy (96.94%), efficient feature extraction, low computational overhead, improved	Dependence on large high-quality datasets, scalability issues in large graphs, limited generalizability,

Author	AI Technology	Distributed Architecture	Application Domain	Benefits	Challenges
	Network (GCN), Transformer Encoder, Multi-head Self-Attention			detection of complex attack patterns, suitable for resource-constrained cloud and IoT systems	computational constraints for model comparisons
(J. Zhou et al., 2023)	Local Search Algorithm, Greedy Search Strategy, Randomized Optimization, Metaheuristic Comparison (GA, PSO)	Hierarchical Distributed Cloud Service Network, Distributed Micro-Clouds and Core Clouds	Smart Cities, IoT Services, Cloud Task Scheduling, 5G-enabled Urban Services	Reduced communication network load, low-latency services, improved cloud efficiency, enhanced data security, better performance than GA and PSO, lower task rejection rate	NP-hard scheduling problems, deadline constraints, resource allocation complexity, avoiding local optima, heterogeneous cloud resources
(Verma and Jailia, 2025)	Deep Learning, Cross-Cloud Threat Transformer (CCTT), Transformer Architecture, LSTM-based IDS, Lemurs Optimizer (LO), Swarm Intelligence	SDN-enabled Multi-Cloud Architecture, Software-Defined Multicloud Defense Controller (SDMDC), Multicloud Intrusion Detection System Gateways (MCIDS-G)	Multi-Cloud Security, Intrusion Detection, Cybersecurity Management, AWS–Azure–Google Cloud Environments	Uniform security enforcement, real-time threat mitigation, centralized policy management, adaptive scalability, automated policy deployment, high detection accuracy, improved resource utilization, strong classification performance	Fragmented security across cloud providers, evasive threats, data breaches, real-time mitigation complexity, heterogeneous traffic patterns, limitations of conventional AI/ML IDS models
(Carrascal et al., 2025)	Machine Learning (ML)-based Inference Services, AI-driven Orchestration, Distributed Intelligence, Data-Driven Decision Making, BentoML Inference Engine	Cloud–Edge–IIoT Continuum, Softwarized Edge Computing, Kubernetes-based Microservices Architecture, SDN-enabled Infrastructure, Containerized Edge-Cloud Platform	Industrial Internet of Things (IIoT), Smart Manufacturing, Industrial Automation, Real-Time Monitoring, Predictive Maintenance, Network Reconfiguration	Low-latency processing, real-time decision making, dynamic network reconfiguration, autonomous adaptation, horizontal autoscaling, efficient CPU and memory utilization, rapid deployment (<5 minutes), sub-second network reconfiguration, scalable and reproducible architecture, improved system responsiveness and operational efficiency	Scalability of industrial deployments, security and privacy requirements, lack of practical end-to-end implementations, integration across cloud-edge-IIoT continuum, maintaining performance under high-density sensor workloads, need for industrial-grade validation datasets and hardware environments
(Ghafir et al., 2024)	Multi-Objective Genetic Algorithm (MOGA), NSGA-II, Bayesian PEFT Ranking, Roulette Selection	Cloud Computing Environment, Virtual Machine (VM)-based Distributed Infrastructure	Scientific Workflow Scheduling, Cloud Resource Management, Pegasus Workflows (LIGO, CyberShake, GENOMIC, SIPHT)	Reduced execution cost (5–6%), reduced time delay (8%), lower energy consumption, improved resource utilization, optimized QoS-aware workflow scheduling	NP-hard scheduling problem, conflicting objectives (cost, energy, deadline, time), resource allocation complexity, multi-objective optimization challenges
(Afrianto et al., 2025)	No direct AI implementation; AI/ML and Deep Learning proposed for future integration,	MVC-based Web Application Architecture, Currently validated in Local Environment (future Distributed Cloud Deployment)	Web Application Security Auditing, Vulnerability Assessment, Public Sector Information Systems	20% faster scanning than OWASP ZAP desktop, 100% detection of high and medium vulnerabilities, reduced alert fatigue, improved	SMTP scalability issues, limited local testing environment, lack of distributed cloud validation, limited low-risk

Author	AI Technology	Distributed Architecture	Application Domain	Benefits	Challenges
	Predictive Threat Modeling			usability (≈80%), actionable remediation recommendations	vulnerability detection coverage
(Poghosyan et al., 2023)	Explainable AI (XAI), Rule-Induction Learning, RIPPER Classification Algorithm, IF-THEN Rule Mining	Cloud-Native Microservices Architecture, Distributed Tracing Framework, Native Cloud Applications	AIOps, Root Cause Analysis (RCA), Cloud Application Monitoring, IT Troubleshooting	Explainable troubleshooting, automated RCA, improved observability, efficient processing of large trace datasets, increased trust in AI recommendations	Complexity of distributed systems, scalability of trace analysis, noisy datasets, overfitting risk, trade-off between explainability and predictive accuracy

Based on Table 1, it can be seen that research on Artificial Intelligence Services in Distributed Cloud environments has shown significant progress in recent years. Various AI technologies have been applied to address various problems in distributed cloud systems, ranging from resource optimization, service scheduling, cybersecurity, system observability, to automated decision-making. Further analysis of these research findings is discussed in the following section to identify technology trends, developments in distributed cloud architectures, ongoing challenges, and future research directions.

DISCUSSION

The Evolution of Artificial Intelligence Technology in Distributed Cloud Environments

The development of Artificial Intelligence (AI) in distributed cloud environments demonstrates a transformation from traditional rules-based and optimization approaches to adaptive, autonomous, and learning-based systems. Broadly speaking, this evolution can be grouped into three main pillars: Optimization-Based AI, Machine Learning-Based AI, and Deep Learning-Based AI. In the early stages, distributed cloud systems relied on optimization methods and decision-support systems such as TOPSIS, VIKOR, PROMETHEE, and hybrid weighting to assist administrators in making resource allocation decisions (Al-Tarawneh et al., 2025). As the network complexity increases, the approach evolves towards a more autonomous mechanism through the integration of Game Theory, Nash Equilibrium, Peak Clipping Algorithm (PCA), local search, greedy search, and Multi-Objective Genetic Algorithm (MOGA) based on NSGA-II which is capable of scheduling and configuring resources independently to maintain quality of service (Quality of Service/QoS) (Xiao et al., 2022; J. Zhou et al., 2023).

The next development is marked by the dominance of Machine Learning (ML) which enables cloud systems to perform predictive analysis, workload forecasting, dynamic resource allocation, and real-time data-driven decision making (Gandhi and Singla, 2026; Priyadarshini et al., 2024). ML implementation also supports cloud-native DevOps automation, including CI/CD processes, service orchestration, and infrastructure management, thereby increasing operational efficiency and reducing human intervention (Nagmoti et al., 2025). In an Edge/Cloud Continuum environment, Carrascal et al. (2025) leveraged the SVM algorithm through BentoML to provide fast inference services on a microservice architecture. However, the limited transparency of ML models prompted the development of Explainable AI (XAI) based on RIPPER, C5.0, and Dempster-Shafer Theory, which is capable of explaining the causes of performance degradation of cloud-native applications through distributed tracing and rule-based reasoning (Poghosyan et al., 2024). In addition, the integration of ML with predictive threat modeling is starting to be used to build more proactive cloud security systems (Afrianto et al., 2025).

At a more advanced stage, Deep Learning (DL) is becoming the dominant technology due to its ability to model complex relationships and high-dimensional data with better accuracy than conventional ML. Architectures like GConvTrans, which combine a Graph Convolutional Network (GCN) and a Transformer Encoder, are able to effectively model spatial and temporal dependencies in distributed cloud data traffic (Appiahene et al., 2026). In the security field, the Cross-Cloud Threat Transformer (CCTT) based on Transformer and LSTM optimized using Lemurs Optimizer shows high capability in detecting cyber attacks across cloud platforms (Verma and Jailia, 2025). In addition, the application of Reinforcement Learning (RL) and in particular Multi-Agent Reinforcement Learning (MARL) allows computational agents to learn autonomously to optimize the allocation of CPU, memory, energy, and resource scheduling based on dynamic environmental conditions (Talhar and Gaikwad, 2025).

Beyond performance improvements and automation, AI also plays a crucial role in security and privacy. AI technology can enhance cloud protection through real-time anomaly detection, predictive threat analysis, and more adaptive cyberattack mitigation (Alakbarov et al., 2026; Ch et al., 2025; Stutz et al., 2024). The use of Reinforcement Learning and Convolutional Neural Networks (CNN) has been proven to increase the accuracy of intrusion detection while reducing the false positive rate (Ch et al., 2025; Selvam and S, 2025). Meanwhile, the concept of distributed intelligence supports adaptive decision making across multiple computing nodes to reduce latency and improve operational efficiency (Myakala, 2025). In contrast, federated learning offers a collaborative learning approach that maintains data privacy by keeping data at its original location (Guberović et al., 2021).

The Development of Distributed Cloud Architecture to Support AI Services

In addition to the evolution of artificial intelligence algorithms, the effectiveness of AI services depends heavily on the development of the underlying cloud computing infrastructure. The evolution of cloud architecture from a centralized (cloud-centric) to a distributed model reflects the industry's need for reduced latency, high availability, and data sovereignty.

In their early stages of development, AI services were largely hosted in centralized (cloud-centric) architectures with massive storage and computing capacities. This design is reflected in the research by Xiao et al. (2022) who designed a centralized dynamic scheduling system, as well as that by Ghafir et al. (2024) who used a Multi-Objective Genetic Algorithm (MOGA) to optimize the scheduling of scientific workloads in a hybrid cloud environment. Although this architecture excels at processing large-scale machine learning models, its reliance on a single data center introduces data transmission latency and a single point of failure. To address these limitations, the architecture is shifting toward load distribution. The concept of a multi-cloud architecture through the Cross-Cloud Threat Transformer (CCTT) (Verma and Jailia, 2025). By deploying an intrusion detection system (IDS) across different cloud platforms, this architecture minimizes the risk of vendor lock-in while enhancing cyber fault tolerance across geopolitical and operational regions.

The next step in the evolution of distributed architecture is the integration of cloud computing with computing units that are closer to end users. Multi-tier architectures such as Edge-Fog-Cloud are increasingly widely adopted. Tactical decision-making using MCDM methods such as TOPSIS and VIKOR can be distributed across the edge, fog, and cloud layers to align computing capacity requirements with strict network latency constraints (Al-Tarawneh et al., 2025). This trend subsequently evolved into the Cloud-Edge-IoT Continuum concept. Contemporary research by Carrascal et al. (2025) demonstrates the full integration of IoT devices, edge nodes, and the central cloud as a seamless cyber-physical system. By providing data-driven inference services directly at the edge, this architecture enables true distributed intelligence, where decisions can be made instantly without sending all raw data back to the cloud data center.

Furthermore, cloud-native technology serves as a vital operational foundation. Microservices architecture has become the de facto standard for building robust and easily scalable AI applications. Poghosyan et al. (2024) emphasize the importance of adopting microservices combined with distributed tracing to detect performance degradation in complex and dynamic systems. Without distributed tracing, isolating errors in long microservice call chains would be extremely difficult for system administrators.

In practice, containerization using Docker and orchestration using Kubernetes have become essential tools for dynamically running AI models. Machine learning inference services can be lightly packaged using inference engines such as BentoML and deployed on microservice container clusters (Carrascal et al., 2025). Orchestration platforms like Kubernetes enable these intelligent agents to automatically scale, self-heal, and transparently migrate tasks across compute nodes to maintain optimal quality of service (QoS) even as workloads fluctuate dramatically.

The Evolution of Distributed Cloud Architecture to Support AI Services

The results of the literature review indicate that the development of Artificial Intelligence (AI) services is inextricably linked to the evolution of distributed cloud architecture, which serves as the primary foundation for providing computing resources. As the complexity of AI models, the volume of data processed, and the need for real-time responses increase, cloud architectures are transforming centralized systems to more distributed, adaptive, and intelligent environments. This transformation is driven by industry demands for reduced latency, improved scalability, energy efficiency, data security, and the ability to manage increasingly dynamic AI workloads.

1. The Evolution of Cloud Architecture from Monolithic to Cloud-Native

The study's findings indicate a shift in cloud architecture from a monolithic model toward a microservices-based, cloud-native architecture to meet the needs of dynamic and scalable AI services. The microservices approach allows each service to be developed, updated, and scaled independently, thus increasing the flexibility of managing AI applications compared to monolithic architectures which tend to be rigid. However, this increased modularity also increases system complexity, necessitating more sophisticated observability mechanisms, such as distributed tracing, to monitor interactions between services and effectively identify root causes. These findings show that cloud evolution is not only focused on scalability, but also on increasingly complex system monitoring and management capabilities. Furthermore, the literature also points to the emergence of serverless computing as a further evolution of cloud-native computing, enabling developers to focus on application logic and AI models without having to directly manage infrastructure. This approach has the potential to accelerate AI service deployment while improving the efficiency of computing resource utilization.

2. Cloud as the Main Foundation for Modern AI Implementation

Literature shows that cloud computing is the main foundation for modern AI implementation because it provides elastic computing resources, large-scale data storage, and high-performance infrastructure for model training and inference processes AI (Al-Tarawneh et al., 2025; Bershchanskyi et al., 2025; Sinha et al., 2026; Zac et al., 2026). Interestingly, the relationship between AI and the cloud is bidirectional. Initially, the cloud served only as an infrastructure provider for AI. Al-Tarawneh et al. (2025) utilized a Multi-Criteria Decision Making (MCDM) approach to optimize task offloading in an Edge-Fog-Cloud environment, while Xiao et al. (2022) used Game Theory to manage workload distribution and reduce peak demand for cloud services. Meanwhile, the use of Multi-Agent Reinforcement Learning (MARL) combined with the LSTM predictive model is able to significantly increase resource utilization, reduce operational costs, and improve compliance with Service Level Agreement (SLA) (Talhar and Gaikwad, 2025; Tidake et al., 2025). These findings indicate the emergence of an AI-Driven Cloud paradigm, a cloud environment capable of resource optimization, workload prediction, and automated decision-making without intensive human intervention. This paradigm provides a critical foundation for the development of next-generation cloud infrastructure.

3. Hybrid Cloud Integration and Edge Computing

Relying solely on cloud data centers is no longer sufficient to meet the needs of modern AI applications that require extremely fast response times. Therefore, the integration of cloud and edge computing has become one of the most prominent trends in the literature (Ganesh et al., 2025; Long et al., 2026). A hybrid cloud–edge approach allows some computational processes to be moved closer to the data source, thereby reducing latency, conserving bandwidth, and improving energy efficiency. Distributing computational processes across the edge, fog, and cloud layers can significantly improve system performance by reducing response times and computational costs (Carrascal et al., 2025; Long et al., 2026). These findings are consistent with the research by Ganesh et al. (2025) who developed a Cloud-Edge-IoT Continuum architecture to support real-time AI inference services in Industrial Internet of Things (IIoT) environments. Thus, the cloud continues to serve as a hub for analytics and large-scale storage, while the edge handles processes that require rapid responses. In other words, the future of AI services is likely to be characterized by cloud-edge synergy rather than the dominance of any single approach.

4. Adoption of Multi-Cloud and Cloud-Agnostic Approaches

The findings of the synthesis also indicate a growing focus on multi-cloud and cloud-agnostic approaches as strategies to avoid vendor lock-in. Reliance on a single cloud provider is seen as potentially posing operational risks, high migration costs, and limited flexibility in managing AI services (Kratzke, 2018; Messina et al., 2017). The multi-cloud concept can be applied to security systems through the use of the Cross-Cloud Threat Transformer (CCTT), which is capable of detecting threats across various cloud platforms in an integrated manner (Verma and Jailia, 2025). In addition to improving interoperability, this approach also strengthens the system's resilience against service failures at a specific cloud provider. From a strategic perspective, these findings indicate that organizations are beginning to view the cloud as an integrated ecosystem of services rather than merely a single infrastructure. Thus, the ability to run AI models across platforms will be a key factor in the future development of distributed cloud.

5. Challenges in Developing a Distributed Cloud for AI

Although it offers significant performance improvements, the implementation of AI in distributed cloud environments still faces various challenges, particularly regarding resource heterogeneity, scheduling complexity, data security, and energy efficiency (Sinha et al., 2026; Zac et al., 2026). From a resource management perspective, the scheduling process in a distributed cloud environment is a complex optimization problem and is often classified as an NP-hard problem (Al-Tarawneh et al., 2025; Xiao et al., 2022; J. Zhou et al., 2023). This complexity increases further when the system must simultaneously take into account various objectives, such as latency, cost, energy consumption, security, and quality of service. From a security perspective, the growing adoption of cloud services has broadened the attack surface (Appiahene et al., 2026; Verma and Jailia, 2025). This situation underscores the need for AI integration to detect cyber threats in real time and in an adaptive manner. In addition, Explainable AI (XAI) enhances the use of AI in infrastructure decision-making, which requires transparency to earn the trust of system administrators (Poghosyan et al., 2023).

The Role of AI in Resource Management Optimization and Task Scheduling

Artificial Intelligence (AI) plays a crucial role in optimizing resource management and task scheduling, particularly in complex environments such as cloud computing and the Internet of Things (IoT). The growing demand for efficient computing resource management is driven by the expansion of IoT devices, AI-based applications, and real-time computing workloads (Andriulo et al., 2024; Liu et al., 2024). AI plays a significant role in improving task scheduling and intelligent resource allocation through algorithms such as Deep Reinforcement Learning (DRL) and metaheuristics. The use of these algorithms has been shown to increase resource utilization by up to 95.36% by ensuring balanced and efficient allocation (Dong and Tuo, 2025), while minimizing resource waste by dynamically adjusting allocations based on real-time needs (Mahalakshmi and Poornima, 2024; Sayiram et al., 2025).

AI-based task scheduling in a cloud environment aims to reduce computation time and lower user costs (J. Zhou et al., 2023). In addition to benefiting users, this approach can also reduce costs for cloud service providers through more sensible management of resource demand and alleviate resource shortages during peak load periods (Xiao et al., 2022). Another important aspect of resource management is task offloading, in which AI is used to dynamically distribute the computational workload to the appropriate resources (Al-Tarawneh et al., 2025). Intelligent task offloading plays a key role in dynamically allocating workloads to maximize resource utilization and maintain quality of service (QoS) in heterogeneous cloud environments (Al-Tarawneh and Saif, 2021; Loutfi et al., 2024). Various AI approaches, such as a combination of Reinforcement Learning–Whale Optimization Algorithm (RL-WOA), Deep Q-Network (DQN), and Artificial Neural Networks (ANN), have been shown to improve scheduling efficiency, optimize task distribution to virtual machines, and significantly reduce task completion time (Kaur et al., 2025; Lin et al., 2020; Zhang et al., 2025). In addition, AI is capable of managing service requests and resolving scheduling issues involving deadline constraints, thereby improving system performance, reducing the load on communication networks, and providing low-latency services in a distributed cloud environment (Urgaonkar et al., 2015).

Various AI and machine learning techniques are widely used. Deep Reinforcement Learning (DRL) has emerged as a reliable paradigm for adaptive and autonomous decision-making in dynamic cloud scheduling (Talhar and Gaikwad, 2025). To address these challenges, a framework that integrates Neuro-Evolutionary Multi-Agent Reinforcement Learning (NE-MARL) enables agents to dynamically develop their own neural policies and coordinate with one another in a decentralized environment (Talhar and Gaikwad, 2025). The concept of Multi-Agent Reinforcement Learning (MARL) is also widely applied in cloud resource allocation and adaptive workload scheduling, particularly in demonstrating a high degree of adaptability for managing tasks in dynamic edge and cloud environments (Lin et al., 2020; Talhar and Gaikwad, 2025; Z. Zhou et al., 2024). In addition, a Long Short-Term Memory (LSTM)-based predictive model is used for workload forecasting, thereby supporting a more proactive, energy-efficient, and cost-effective allocation of resources (Prasanth et al., 2025; Zheng and Jiaqing, 2025).

On the other hand, a hybrid framework that combines machine learning and heuristic scheduling has been shown to improve task allocation and security-aware resource provisioning, while the Whale Optimization Algorithm (WOA) is capable of optimizing energy consumption without compromising quality of service (QoS) (Goyal et al., 2021). Hybrid swarm intelligence approaches, including the Antlion Optimization Algorithm, are also used to improve the efficiency and adaptability of multi-objective task scheduling (Abualigah and Diabat, 2021). However, traditional swarm algorithms have high computational costs and scalability limitations in large-scale distributed environments. As an alternative, several local search algorithms have demonstrated better performance than Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) in solving task scheduling problems (J. Zhou et al., 2023). In addition, evolutionary algorithms such as GA and the Artificial Flora Scheduler are capable of producing more optimal solutions with shorter execution times (Bacanin et al., 2019; Sayiram et al., 2025), Meanwhile, deep learning is widely used to address data imbalance and cloud complexity, thereby improving the accuracy, efficiency, and scalability of systems (Lin et al., 2020; Wang et al., 2024).

The implementation of AI in cloud resource management also has a significant positive impact on environmental and economic sustainability. AI directly contributes to energy efficiency in cloud data centers; for example, DRL-based systems can reduce energy consumption by up to 10% compared to traditional methods (Dong and Tuo, 2025; Lin et al., 2020). From a financial perspective, the use of AI-based predictions and process automation enables organizations to significantly reduce operating costs through more efficient resource management (Benchrif et al., 2026; Winata et al., 2025). Despite these technological advances, managing cloud resources remains a challenge, particularly when it comes to optimizing scheduling workflows (Ghafir et al., 2024). Many of the scheduling algorithms currently used in cloud environments often focus solely on server availability or idle time, while ignoring the temporal attributes of the cloud tasks themselves (Xiao et al., 2022).

The Role of AI in Distributed Cloud System Security

Artificial Intelligence (AI) plays a crucial role in strengthening security systems in distributed and multi-cloud environments by providing advanced capabilities for threat detection and mitigation. The highly dynamic, heterogeneous, and complex nature of distributed systems demands intelligent solutions to protect data integrity, application reliability, and the resilience of the entire infrastructure from cyber threats that continue to evolve non-deterministically (Sathupadi, 2022; Seth et al., 2024; Verma and Jailia, 2025). As a concrete example, the integration of Software-Defined Networking (SDN) technology combined with deep learning algorithms has been proven to facilitate adaptive attack detection and dynamic security policy management to address operational complexity and network traffic fluctuations in multi-cloud architectures (Verma and Jailia, 2025).

In the realm of threat detection, the use of Machine Learning (ML) and Deep Learning (DL) techniques has been widely implemented to identify various types of cyberattacks with a high degree of precision. AI-based

Network Intrusion Detection Systems (NIDS) are a crucial pillar for flagging suspicious activity on cloud infrastructure (Appiahene et al., 2026; Farooq et al., 2019). The application of Artificial Neural Networks (ANN) is very effective in predicting and detecting security breaches in real-time, including improving the accuracy of botnet detection in cloud environments by analyzing device behavior anomalies (Kanimozhi and Jacob, 2021). Meanwhile, DL techniques are integrated into an intelligent hybrid IDS system for real-time anomaly analysis in large-scale environments, while improving detection performance and reducing model training time in multi-cloud IoT ecosystems (Verma and Jailia, 2025). For more specific attacks such as Distributed Denial of Service (DDoS), the use of Convolutional Neural Networks (CNN) has been proven to provide superior detection results compared to other conventional models due to CNN's ability to recognize spatial and volumetric patterns of suspicious data packets (Verma et al., 2024). In addition, Long Short-Term Memory (LSTM) based models are very reliable in tracking temporal attack patterns that occur gradually and slowly, such as slow-rate DDoS, progressive brute force attacks, and gradual data breaches that often escape traditional threshold detection (Verma and Jailia, 2025). The use of layered neural networks such as in the Multi-User Security Enhancement (MUSE) system is also effective in identifying malicious activities such as data payload manipulation in the data flow between the cloud core and the IoT gateway (Gupta et al., 2022), Although anomaly detection methods in multi-cloud environments are generally still a subject of ongoing research.

Next-generation AI models play a critical role in enhancing distributed cloud security through more adaptive, standardized, and scalable protection mechanisms across multi-cloud environments (Seth et al., 2024). Integrating AI models trained at the edge into the cloud, such as in the MUSE architecture, has been shown to reduce training time while improving local detection accuracy (Gupta et al., 2022). In addition, the combination of a lightweight Intrusion Detection System (IDS) and AI-based policy management improves resource utilization efficiency and security system scalability (Verma and Jailia, 2025). The development of solutions such as EOS-IDS also demonstrates improved threat detection capabilities by reducing the risk of overfitting and false positives commonly found in conventional IDS (Soussi et al., 2023).

Despite offering significant security improvements, implementing AI in distributed clouds still faces various challenges, such as the reliance on large, high-quality attack datasets to generate accurate detection models. Furthermore, AI/ML models often struggle to capture complex dynamic patterns and non-linear relationships in heterogeneous cloud environments, potentially increasing the false positive rate and reducing the system's generalizability (Boryło et al., 2024). Further research needs to expand the focus not only on intrusion detection, but also on autonomous data encryption, firewall configuration automation, secure network traffic management, and detection of advanced attacks such as zero-day attacks.

The Role of Explainable AI (XAI) in Enhancing Trust in Cloud Systems

In the modern IT operations ecosystem or AI for IT Operations (AIOps), transparency and interpretability of Machine Learning (ML)-based decision-making algorithms have become a crucial need (Arrieta, 2020). Given that AIOps systems process massive volumes of data to automate cloud infrastructure operations, stakeholders absolutely need a deep understanding of how these decisions are made in order to build trust and facilitate harmonious collaboration between humans and AI (Poghosyan et al., 2024). In practical industrial contexts, explainable outcomes are often valued far more than highly accurate predictive models that operate as opaque black boxes; therefore, rule learners are preferred when simplicity and human interpretability are valued more than mere predictive power.

To address the opacity challenge, rule-induction systems such as Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and C5.0 rule-based classification emerge as powerful XAI solutions as they offer an optimal balance between logical explanation and decision-making efficacy (Poghosyan et al., 2024; Quinlan, 2014). This rule induction classification is capable of producing deductive logic rules that are inherently easy to understand by system administrators, particularly in facilitating root-cause analysis (RCA). Through this transparent logical structure, humans can easily trace the location of system failures under specific conditions, thereby speeding up the debugging process, performance optimization, and formulation of preventive actions in the future. Furthermore, the operational efficiency of this rule learning system can be dramatically improved through a sampling strategy that has been proven to reduce program execution time especially on noisy datasets while reducing the length of recommendations so that administrators can focus on the most critical anomalies.

Although rule induction systems offer high interpretability, the uncertainty aspect in dynamic cloud data streams remains a challenge that needs to be validated. To mitigate this, a Dempster-Shafer Evidence Theory (DST)-based inference framework can be integrated to verify the uncertainty level of recommendations generated by the rule learning model (Poghosyan et al., 2024). DST allows the combination of multiple independent knowledge sources for robust reasoning and can be used to build classification methods that support user-defined rule validation. Through the what-if analysis framework provided by DST, system engineers can gain a more comprehensive understanding of the behavior of underlying running cloud applications.

Despite offering a promising technological leap, the application of rule induction-based XAI in cloud environments still leaves significant research gaps. Future research should explore the use of alternative ML models beyond traditional rule induction and assess how varying sampling strategies affect their predictive performance (Cohen, 1995). In addition, in-depth investigations into the refinement of the sampling process, the scalability of strategies in handling large-scale datasets on complex cloud infrastructures, and the formulation of specific evaluation metrics to measure the impact of sampling on model accuracy, efficiency, and interpretability are absolutely necessary. Through comparative studies with other classification techniques and expanding the scope into the realm of predictive analysis, anomaly detection, and decision support systems, the direction of XAI technology development is expected to provide a much safer, autonomous, and transparent cloud monitoring posture.

AI-Powered AIOps and Autonomous Cloud Management

1. The Urgency of Transition from Manual to Autonomous Operations

The highly distributed and dynamic landscape of modern cloud infrastructure has surpassed the capacity of human manual management. System administrators and Site Reliability Engineers (SREs) are no longer able to make real-time decisions due to the massive and complex complexity of background processes (Poghosyan et al., 2024). Traditional logging methods are considered inadequate to handle the volume and complexity of data generated by microservice-based architectures capable of elastic resource allocation. The limitations of conventional heuristic and metaheuristic approaches in adapting to dynamic workload fluctuations have led to the emergence of artificial intelligence-based cloud management solutions (Talhar and Gaikwad, 2025). The synergy between AIOps (AI for IT Operations) and autonomous management is present to transform information technology (IT) operations from being reactive and manual into a proactive system that is capable of developing independently (self-evolving).

2. AI-Powered AIOps Paradigm: Observability and Trust

AIOps relies on machine learning-based software intelligence to automate the development cycle, service delivery, and troubleshooting (Cohen, 1995; Poghosyan et al., 2024). One of its key pillars is distributed tracing, which provides continuous observability of request flows across microservices to instantly detect performance bottlenecks. When combined with Explainable AI (XAI), AIOps can automate Root Cause Analysis (RCA) while increasing the transparency of the AI decision-making process. This approach helps reduce the risk of bias, increase user trust, and support more effective collaboration between humans and AI systems. In addition to accelerating the time to detect and remediate cyber intrusions (reducing the time to data entry (MTTD) and time to data entry (MTTR)), AIOps also reduces data storage operational costs through an intelligent sampling strategy that selects only the most relevant subset of trace data.

3. Autonomous Cloud Management: Decentralization, Optimization, and Prediction

While AIOps focuses on observability and problem diagnosis, Autonomous Cloud Management goes further into the realm of autonomous control through dynamic optimization of resource allocation. Cutting-edge frameworks like MARL-DPMRP integrate Multi-Agent Reinforcement Learning (MAL) with LSTM-based deep predictive models to deliver environmentally conscious resource management (Talhar and Gaikwad, 2025). In this decentralized architecture, each Virtual Machine (VM) acts as an autonomous agent that learns local allocation policies, while centralized training guides the agent toward the optimal balance between performance, cost, and sustainability. The integration of LSTM-based workload prediction allows the system to take preventive action before demand spikes occur.

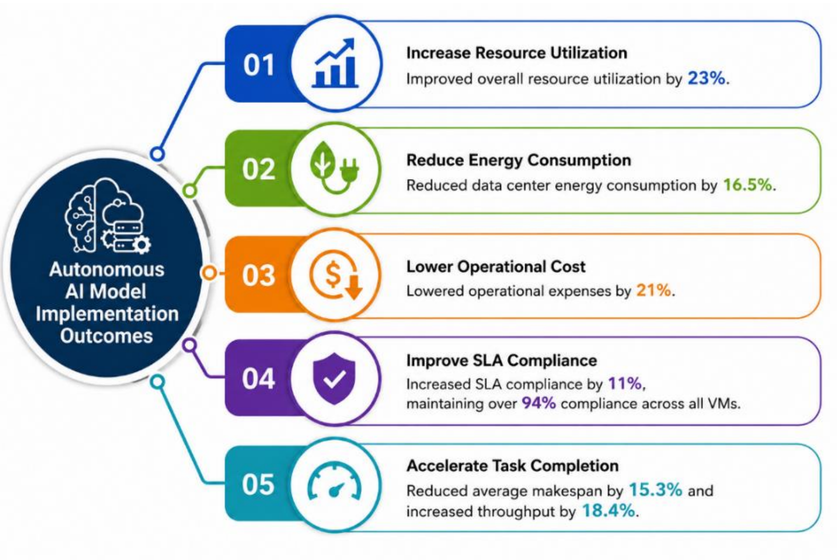

Figure 3. Implementation of autonomous model

4. Implementation on Cloud-Edge-IoT Continuum and Kubernetes Orchestration

These autonomous systems find particularly strong relevance when applied to modern industrial ecosystems such as the Industrial Internet of Things (IIoT) that operate along the cloud/edge/IoT continuum (Carrascal et al., 2025). The use of a container-based microservices architecture enables the infrastructure to take advantage of built-in capabilities such as auto-scaling, load balancing, and fault tolerance. Adaptive decisions and dynamic network reconfiguration are facilitated by a control plane that communicates RESTfully with an inference engine (such as BentoML) in the cloud. This decision loop, which combines real-time telemetry, ML inference, and autonomous policy enforcement, enables instantaneous detection of anomalous behavior.

Performance evaluations show that this Kubernetes-based system is highly efficient, with a fully operational initialization time of under 270 seconds. Furthermore, this continuum architecture offers extremely low latency, with the BentoML inference service maintaining an average response time below 350 ms, InfluxDB database read latency consistently below 500 ms, and estimated network reconfiguration time of only around 800 ms. While this sub-second soft real-time performance does not yet meet the millisecond determinism demands of hard real-time industrial-scale control loops, the system is highly reliable for supporting autonomous applications such as predictive maintenance and anomaly monitoring where elastic scaling is crucial for maintaining service availability.

Key Challenges of Artificial Intelligence Implementation in Distributed Cloud Environments

1. Complexity of Cyber Security and Threat Detection

Implementing AI in distributed and multi-cloud environments faces security challenges due to infrastructure heterogeneity and differences in security mechanisms across service providers, which increases management complexity and expands the potential for cyberattacks. Therefore, adaptive and efficient AI models are needed to maintain the security of data, applications, and cloud infrastructure. However, Machine Learning and Deep Learning-based IDS still face limitations in generalizing across cloud environments, making them prone to producing false positives and less adaptive to changes in network traffic. Model performance also depends heavily on the availability of large, high-quality attack datasets and the ability to detect non-linear anomaly patterns in distributed environments. While AI offers real-time threat detection and dynamic security policies, its long-term effectiveness still requires ongoing optimization and evaluation to operate optimally in multi-cloud environments.

2. Optimization of Resource Management and Task Scheduling Complexity

Efficient resource allocation in distributed clouds is a major challenge due to dynamic workloads, demand fluctuations, and energy consumption constraints. Traditional optimization methods and metaheuristics such as First Fit, PSO, and ACO are considered less adaptive, have slow convergence, and are less scalable for modern cloud environments. The use of swarm intelligence techniques on large-scale heterogeneous infrastructure also incurs high computational overhead. Meanwhile, multi-objective genetic algorithms like SPEA2 still face limitations in convergence time and execution efficiency for real-time decision-making needs. Furthermore, the task scheduling problem in distributed clouds is NP-hard, making finding optimal solutions increasingly complex as the number of tasks increases. Consequently, modern cloud systems still face challenges in coordinating decentralized learning, multi-agent interactions, and integrating optimization objectives with energy efficiency and environmental sustainability.

3. Observability and Transparency of AI Decisions Barriers

Disturbance diagnosis in distributed clouds is becoming increasingly complex due to non-deterministic interdependencies between components. Therefore, AI-based monitoring systems are used to support early detection, Root Cause Analysis (RCA), and rapid performance recovery. However, the increasing complexity of cloud infrastructure makes real-time decision-making by administrators increasingly difficult, necessitating an intelligent, accurate and transparent RCA system through Explainable AI (XAI). Transparency of AI decisions is also crucial in multi-tenant environments to increase user trust and ensure analysis results can be verified by human experts.

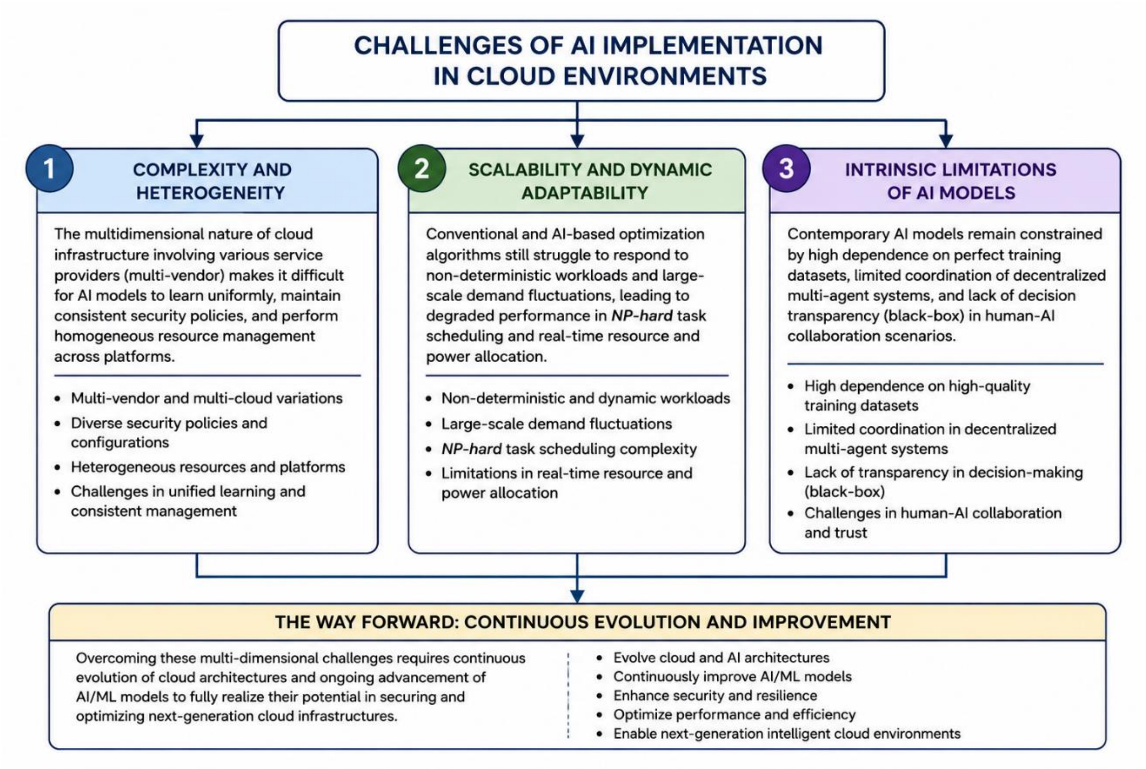

Figure 4. Synthesis of Dominant Challenge Categories

Future Research Directions in AI-Based Distributed Cloud Computing

Future research on Artificial Intelligence (AI)-based distributed cloud is expected to focus on developing a more intelligent, secure, efficient, and autonomous cloud ecosystem. One of the main priorities is improving the explainability of AI models to strengthen user trust and support more effective collaboration between humans and intelligent systems in managing complex cloud environments, particularly in high-risk multi-tenant infrastructures. In addition, challenges related to decentralized learning and coordination of multi-agent systems remain important research areas to realize fully autonomous resource management without reliance on centralized control. At the same time, the integration of sustainability and energy efficiency metrics into cloud resource allocation mechanisms is gaining increasing attention as part of sustainable infrastructure development strategies. Through a combination of multi-agent optimization, predictive modeling, and energy-aware resource management, the future evolution of the cloud is projected to lead to intelligent resource provisioning that can optimize performance while minimizing energy consumption. This development will also support the implementation of adaptive workload scheduling that is more dynamic and responsive to changes in workloads in increasingly complex distributed cloud environments.

On the other hand, strengthening the cybersecurity sector and anomaly detection in distributed cloud environments demands continuous evolution and refinement of contemporary AI/ML models. Although AI-based anomaly detection has shown promising potential on single-cloud infrastructures, its application to heterogeneous multi-cloud environments is still in the early stages of research (nascent study area) and requires more in-depth exploration. Future research agendas should focus on addressing the traditional limitations found in existing AI/ML-based Intrusion Detection Systems (IDS), particularly the failure of models to effectively generalize learning across multiple cloud environments, excessive false positive rates that burden administrators, and the lack of model adaptability to fluctuating and noisy changes in network traffic patterns.

To bridge this gap, the integration of intelligent models into next-generation cloud network architectures is becoming one of the most promising areas of development. Complex modern cloud architectures can significantly benefit from real-time threat mitigation and dynamic policy enforcement capabilities, which have been empirically proven to outperform existing conventional multi-cloud security methods. However, the integration of AI-based threat detection models into multi-cloud ecosystems based on Software-Defined Networking (SDN) is currently far from perfect and often experiences performance degradation. Therefore, refining the functional integration between a flexible SDN control plane and an adaptive AI inference engine is a crucial research direction for creating a resilient, adaptive, and autonomous cyber defense posture in future cloud infrastructures.

CONCLUSION

This study reviews the application of Artificial Intelligence (AI) in distributed cloud environments through an analysis of 11 recent studies selected from Scopus-indexed publications. The findings indicate that AI has become a key driver for improving resource management, workload scheduling, security, service optimization, and autonomous decision-making in distributed cloud infrastructures. Recent developments demonstrate a clear transition from traditional optimization approaches to techniques based on Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL), enabling adaptive, predictive, and autonomous cloud operations.

The literature further reveals that distributed cloud architectures are evolving from conventional monolithic systems to cloud-native, microservices-based, edge-cloud, and multi-cloud environments. This transformation has increased the need for intelligent orchestration mechanisms capable of managing heterogeneous resources, dynamic workloads, and complex service interactions. AI-based approaches, including Explainable AI (XAI), federated learning, graph-based deep learning, and multi-agent reinforcement learning, have shown significant potential in addressing these challenges while improving operational efficiency, security, scalability, and Quality of Service (QoS).

Despite this progress, several challenges remain. The reviewed studies highlight persistent issues related to infrastructure heterogeneity, large-scale resource optimization, cybersecurity threats, the explainability of AI decisions, and the limited generalizability of current AI models across different cloud environments. These limitations suggest that further research is needed to develop more robust, transparent, and adaptive AI-based cloud management frameworks.

Overall, evidence suggests that distributed cloud computing is evolving toward a highly autonomous ecosystem characterized by intelligent resource provisioning, self-optimization, self-healing capabilities, and adaptive security mechanisms. Future research is expected to focus on AI-based cloud management, federated and distributed learning, edge intelligence, continuous cloud computing, and trustworthy AI, positioning AI-powered distributed cloud infrastructure as a fundamental foundation for next-generation digital services.

REFERENCES

Abdullah, P. Y., Zeebaree, S. R. M., Shukur, H. M., & Jacksi, K. (2020). HRM system using cloud computing for Small and Medium Enterprises (SMEs). *Technology Reports of Kansai University, 62*(04), 1977–1987. https://www.researchgate.net/profile/Subhi-Zeebaree/publication/341883552_HRM_System_using_Cloud_Computing_for_Small_and_Medium_Enterprises_SMEs/links/5ed829f992851c9c5e7b7d27/HRM-System-using-Cloud-Computing-for-Small-and-Medium-Enterprises-SMEs.pdf

Abdulqader, D. M., & Zeebaree, S. R. M. (2021). Impact of Distributed-Memory Parallel Processing Approach on Performance Enhancing of Multicomputer-Multicore Systems A Review. *Qalaai Zanist Scientific Journal, 6*(4), 1134–1140. https://doi.org/10.25212/lfu.qzj.6.4.46

Abualigah, L., & Diabat, A. (2021). A novel hybrid antlion optimization algorithm for multi-objective task scheduling problems in cloud computing environments. *Cluster Computing, 24*(1), 205–223. https://doi.org/10.1007/s10586-020-03075-5

Afrianto, I., Setyaningrum, I., Atin, S., Nugroho, E. P., & Digdo, G. P. (2025). Security Engineering of Open Web Application Security Project-Based Automated System for Web Vulnerability Detection and Mitigation. *International Journal of Safety and Security Engineering, 15*(12). https://doi.org/10.18280/ijsse.151206

Al-Tarawneh, M. A. B., Kanj, H., & Aly, W. H. F. (2025). An integrated MCDM framework for trust-aware and fair task offloading in heterogeneous multi-provider Edge-Fog-Cloud systems. *Results in Engineering, 26*, 105228. https://doi.org/10.1016/j.rineng.2025.105228

Al-Tarawneh, M. A. B., & Saif. (2021). Performance Assessment of Context-aware Online Learning for Task Offloading in Vehicular Edge Computing Systems. *International Journal of Advanced Computer Science and Applications, 12*(4). https://doi.org/10.14569/IJACSA.2021.0120439

Alakbarov, R., Hashimov, M., & Alakbarov, O. (2026). Artificial Intelligence based Behavior Adaptive Authentication Model in Solving Security Problems of Cloud Computing Systems. *2026 6th International Conference on Expert Clouds and Applications (ICOECA)*, 781–785. https://doi.org/10.1109/ICOECA68095.2026.11485201

Andriulo, F. C., Fiore, M., Mongiello, M., Traversa, E., & Zizzo, V. (2024). Edge Computing and Cloud Computing for Internet of Things: A Review. *Informatics, 11*(4), 71. https://doi.org/10.3390/informatics11040071

Appiahene, P., Berchie, S. O., Botchway, E., Ayitey, M. J., Dawson, J. K., Mettle, H. N.-A., & Afrifa, S. (2026). Network intrusion detection using a hybrid graph-based convolutional network and transformer architecture. *PLOS One, 21*(1), e0340997. https://doi.org/10.1371/journal.pone.0340997

Arrieta, A. B. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bacanin, N., Tuba, E., Bezdan, T., Strumberger, I., & Tuba, M. (2019). Artificial Flora Optimization Algorithm for Task Scheduling in Cloud Computing Environment. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 437–445). https://doi.org/10.1007/978-3-030-33607-3_47

Benchrif, M., Jamali, A., & Naja, N. (2026). Toward Sustainable Cloud Computing: A Multi-Layer Taxonomy of AI-Based Resource Optimization and Scheduling Methods. *In Lecture Notes in Networks and Systems* (pp. 114–121). https://doi.org/10.1007/978-3-032-19760-3_13

Bershchanskyi, Y., Oberyshyn, R., & Klym, H. (2025). Designing Scalable Architectures for Cloud-Based AI-Driven Systems. *2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 1016–1020. https://doi.org/10.1109/IDAACS68557.2025.11322157

Boryło, P., Chołda, P., Domżał, J., Jaglarz, P., Jurkiewicz, P., Rzepka, M., Rzym, G., & Wójcik, R. (2024). SDNRoute: Proactive routing optimization in Software Defined Networks. *Computer Communications, 225*, 250–278. https://doi.org/10.1016/j.comcom.2024.07.015

Carrascal, D., Díaz-Fuentes, J., Manso, N., Lopez-Pajares, D., Rojas, E., Savi, M., & Arco, J. M. (2025). Softwarized Edge Intelligence for Advanced IIoT Ecosystems: A Data-Driven Architecture Across the Cloud/Edge Continuum. *Applied Sciences, 15*(19), 10829. https://doi.org/10.3390/app151910829

Ch, R., Nimmala, S., Batra, I., Malik, A., & Malik, P. K. (2025). Enhancing Cloud Security and Efficiency Through AI-Driven Intrusion Detection and Machine Learning-Based Resource Management. *In Deep Learning Innovations for Securing Critical Infrastructures* (pp. 241–256). https://doi.org/10.4018/979-8-3373-0563-9.ch015

Cohen, W. W. (1995). Fast Effective Rule Induction. *In Machine Learning Proceedings 1995* (pp. 115–123). Elsevier. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

Dong, R., & Tuo, M. (2025). Research on cloud computing resource optimization scheduling based on artificial intelligence. In J. Shi, R. Damaševičius, S.-K. Chang, & P. Geng (Eds.), International Conference on Optoelectronics and Information Technology (ICOIT 2025) (p. 107). *SPIE*. https://doi.org/10.1117/12.3071304

Duan, S., Wang, D., Ren, J., Lyu, F., Zhang, Y., Wu, H., & Shen, X. (2023). Distributed Artificial Intelligence Empowered by End-Edge-Cloud Computing: A Survey. *IEEE Communications Surveys & Tutorials, 25*(1), 591–624. https://doi.org/10.1109/COMST.2022.3218527

Farooq, M. S., Riaz, S., Abid, A., Abid, K., & Naeem, M. A. (2019). A Survey on the Role of IoT in Agriculture for the Implementation of Smart Farming. *IEEE Access, 7*, 156237–156271. https://doi.org/10.1109/ACCESS.2019.2949703

Gandhi, R., & Singla, J. (2026). A Systematic Review of Artificial Intelligence Techniques for Distributed Storage and Data Processing in Cloud Computing. *2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL),* 573–578. https://doi.org/10.1109/ICSADL67539.2026.11452031

Ganesh, G., Nasreen, M. A., & Jaganathan, A. (2025). Hybrid Edge-Cloud AI and Blockchain System Architecture for Real-Time Decision-Making in IT Project Management. *2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS),* 751–755. https://doi.org/10.1109/ICACRS67045.2025.11324399

Ghafir, S., Alam, M. A., Siddiqui, F., Naaz, S., Sohail, S. S., & Madsen, D. Ø. (2024). Toward optimizing scientific workflow using multi-objective optimization in a cloud environment. *Cogent Engineering, 11*(1). https://doi.org/10.1080/23311916.2023.2287303

Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., Golec, M., Stankovski, V., Wu, H., Abraham, A., Singh, M., Mehta, H., Ghosh, S. K., Baker, T., Parlikad, A. K., Lutfiyya, H., Kanhere, S. S., Sakellariou, R., Dustdar, S., … Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things, 19,* 100514. https://doi.org/10.1016/j.iot.2022.100514

Goyal, S., Bhushan, S., Kumar, Y., Rana, A. ul H. S., Bhutta, M. R., Ijaz, M. F., & Son, Y. (2021). An Optimized Framework for Energy-Resource Allocation in a Cloud Environment based on the Whale Optimization Algorithm. *Sensors, 21*(5), 1583. https://doi.org/10.3390/s21051583

Guberović, E., Lipić, T., & Čavrak, I. (2021). Dew Intelligence: Federated learning perspective. *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1819–1824.

https://doi.org/10.1109/COMPSAC51774.2021.00274

Gupta, L., Salman, T., Ghubaish, A., Unal, D., Al-Ali, A. K., & Jain, R. (2022). Cybersecurity of multi-cloud healthcare systems: A hierarchical deep learning approach. *Applied Soft Computing, 118*, 108439. https://doi.org/10.1016/j.asoc.2022.108439

Kanimozhi, V., & Jacob, T. P. (2021). Artificial Intelligence outflanks all other machine learning classifiers in Network Intrusion Detection System on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. *ICT Express, 7*(3), 366–370. https://doi.org/10.1016/j.icte.2020.12.004

Kaur, S., Singh, J., & Bharti, V. (2025). An optimised AI-driven swarm-based enhanced task scheduling model for cloud computing environment. *International Journal of Cloud Computing, 14*(1), 25–53. https://doi.org/10.1504/IJCC.2025.145660

Kratzke, N. (2018). A Brief History of Cloud Application Architectures. *Applied Sciences, 8*(8), 1368. https://doi.org/10.3390/app8081368

Kunduru, A. R. (2023). Artificial intelligence usage in cloud application performance improvement. *Central Asian Journal of Mathematical Theory and Computer Sciences, 4*(8), 42–47. https://cajmtcs.casjournal.org/index.php/CAJMTCS/article/view/491

Lin, J., Cui, D., Peng, Z., Li, Q., & He, J. (2020). A Two-Stage Framework for the Multi-User Multi-Data Center Job Scheduling and Resource Allocation. *IEEE Access, 8*, 197863–197874. https://doi.org/10.1109/ACCESS.2020.3033557

Liu, X., Yu, P., & Liu, X. (2024). Cloud Computing Task Scheduling based on Improved Particle Swarm Algorithm. Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering, 198–202. https://doi.org/10.1145/3689236.3691495

Long, N. Van, Vinh, T. N., Quyen, H. Van, Duyen, N. T. M., Hoa, P. Van, & Tuan, M. A. (2026). The Evolution of AI Chips: From Cloud to Edge, Powering the Future of Artificial Intelligence. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST,* 188–217. https://doi.org/10.1007/978-3-032-14055-5_14

Loutfi, S. I., Shayea, I., Tureli, U., El-Saleh, A. A., & Tashan, W. (2024). An overview of mobility awareness with mobile edge computing over 6G network: Challenges and future research directions. *Results in Engineering, 23,* 102601. https://doi.org/10.1016/j.rineng.2024.102601

Mahalakshmi, V., & Poornima, V. (2024). Multi-Objective Task Scheduling in the Cloud Environment Using Reinforcement Learning and Whale Optimization Algorithm. *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA),* 501–506. https://doi.org/10.1109/ICSCNA63714.2024.10863878

Messina, F., Mikkilineni, R., Morana, G., & Rosaci, D. (2017). Track Chair's Report: Convergence of Distributed Clouds, Grids and their Management CDCGM 2017. *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE),* 92–94. https://doi.org/10.1109/WETICE.2017.34

Myakala, P. K. (2025). Distributed Data Intelligence in the Edge-to-Cloud Continuum: a Survey of Architectures, Models, and Applications. *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, 93–99. https://doi.org/10.1109/ICDICI66477.2025.11135068

Nagmoti, N. S., Srivastava, I., & Damle, M. (2025). AI-Driven Enhancements in Cloud-Native DevOps Boosting Automation, Deployment, and Monitoring. *In Artificial Intelligence for Cloud-Native Software Engineering* (pp. 203–236). https://doi.org/10.4018/979-8-3693-9356-7.ch008

Poghosyan, A., Harutyunyan, A., Davtyan, E., Petrosyan, K., & Baloian, N. (2024). The Diagnosis-Effective Sampling of Application Traces. *Applied Sciences, 14*(13), 5779. https://doi.org/10.3390/app14135779

Poghosyan, A., Harutyunyan, A., Grigoryan, N., & Pang, C. (2023). Distributed Tracing for Troubleshooting of Native Cloud Applications via Rule-Induction Systems. *JUCS - Journal of Universal Computer Science, 29*(11), 1274–1297. https://doi.org/10.3897/jucs.112513

Prasanth, A., Babu, K. N., Rahul, P., & Reddy, B. V. (2025). A Deep Learning-Based Workload Forecasting Model in Cloud Data Centers. *2025 International Conference on Computing Technologies (ICOCT),* 1–5. https://doi.org/10.1109/ICOCT64433.2025.11118402

Priyadarshini, S., Sawant, T. N., Bhimrao Yadav, G., Premalatha, J., & Pawar, S. R. (2024). Enhancing security and scalability by AI/ML workload optimization in the cloud. *Cluster Computing, 27*(10), 13455–13469. https://doi.org/10.1007/s10586-024-04641-x

Quinlan, J. R. (2014). C4.5: Programs for Machine Learning. *Elsevier*.

Sandhu, I. K., Malhotra, M., & Randhawa, P. R. (2022). Enhancing the Secured Automation of Storage process on Cloud: Embedding the features of Artificial Intelligence and AI-powered Network Security Tools. *International Journal of Early Childhood Special Education, 14*(5).

Sathupadi, K. (2022). AI-driven QoS optimization in multi-cloud environments: investigating the use of AI techniques to optimize QoS parameters dynamically across multiple cloud providers. *Applied Research in Artificial Intelligence and Cloud Computing, 5*(1), 213–226.

Sayiram, S., Vedula, J., Baskaran, S., Muthusamy, R., Jiwani, N., & Kiruthiga, T. (2025). Optimizing Real-Time Task Scheduling in Cloud-based AI Systems using Genetic Algorithms. *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS),* 1–5. https://doi.org/10.1109/ICICACS65178.2025.10967710

Selvam, M., & S, K. B. (2025). AI-Powered Cloud Computing for Performance Optimization and Scalability in Distributed Systems. *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF),* 1–6. https://doi.org/10.1109/COMP-SIF65618.2025.10969950

Seth, D., Nerella, H., Najana, M., & Tabbassum, A. (2024). Navigating the Multi-Cloud Maze: Benefits, Challenges, and Future Trends. *International Journal of Global Innovations and Solutions (IJGIS)*. https://doi.org/10.21428/e90189c8.8c704fe4

Sinha, A., Rana, M. E., Hameed, V. A., & Al-Othmani, A. Z. (2026). A Catalyst for Artificial Intelligence and Machine Learning Revolution: Exploring Synergies, Challenges, and the Future of AI/ML in the Cloud. *Lecture Notes in Networks and Systems*, 501–515. https://doi.org/10.1007/978-981-96-9245-3_39

Soussi, W., Christopoulou, M., Gür, G., & Stiller, B. (2023). MERLINS – Moving Target Defense Enhanced with Deep-RL for NFV In-Depth Security. *2023 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN),* 65–71. https://doi.org/10.1109/NFV-SDN59219.2023.10329594

Stutz, D., de Assis, J. T., Laghari, A. A., Khan, A. A., Andreopoulos, N., Terziev, A., Deshpande, A., Kulkarni, D., & Grata, E. G. H. (2024). Enhancing Security in Cloud Computing Using Artificial Intelligence. *In Applying Artificial Intelligence in Cybersecurity Analytics and Cyber Threat Detection* (pp. 179–220). Wiley. https://doi.org/10.1002/9781394196470.ch11

Talhar, N. R., & Gaikwad. (2025). Intelligent Cloud Resource Provisioning Using Multi-agent Reinforcement Learning and Deep Predictive Modelling. *International Journal of Intelligent Engineering and Systems, 19*(1), 880–895. https://doi.org/10.22266/ijies2026.0131.53

Tidake, S. P., Mulkalwar, P. N., & Atique, M. (2025). Intelligent Resource Allocation Strategies in Cloud Computing: A Review and Research Outlook. *16th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2025*.

Urgaonkar, R., Wang, S., He, T., Zafer, M., Chan, K., & Leung, K. K. (2015). Dynamic service migration and workload scheduling in edge-clouds. *Performance Evaluation, 91*, 205–228. https://doi.org/10.1016/j.peva.2015.06.013

Verma, R., & Jailia, M. (2025). SDN-enabled adaptive security framework for multi-cloud infrastructures using deep learning-based threat detection and policy management. *PeerJ Computer Science, 11*, e3266. https://doi.org/10.7717/peerj-cs.3266

Verma, R., Jailia, M., Kumar, M., & Kaliraman, B. (2024). CNN-Based Detection of DDoS Attacks in Multi-Cloud Environments. *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, 1011–1016. https://doi.org/10.1109/IC3SE62002.2024.10593351

Wang, C., Guo, R., Xiu, P., Li, X., & Wang, H. (2024). Research on Resource Scheduling Algorithm Optimization in Cloud Computing Environment Based on Deep Learning. *2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC),* 918–922. https://doi.org/10.1109/EIECC64539.2024.10929098

Winata, K. C., Putra, J. D. E., Setiawan, R., & Rumagit, R. Y. (2025). A Literature Review of AI-Driven Cloud Computing. *2025 International Conference on ICT for Smart Society (ICISS),* 1–8. https://doi.org/10.1109/ICISS66954.2025.11389645

Wu, Y. (2020). Cloud-edge orchestration for the Internet of Things: Architecture and AI-powered data processing. *IEEE Internet of Things Journal, 8*(16), 12792–12805. https://ieeexplore.ieee.org/abstract/document/9162084

Xiao, Z., Wang, M., Chronopoulos, A. T., & Jiang, J. (2022). A method for reducing cloud service request peaks based on game theory. *Journal of Parallel and Distributed Computing, 165*, 107–119.

https://doi.org/10.1016/j.jpdc.2022.03.002

Xu, Z., Liu, W., Huang, J., Yang, C., Lu, J., & Tan, H. (2020). Artificial Intelligence for Securing IoT Services in Edge Computing: A Survey. *Security and Communication Networks, 2020*, 1–13. https://doi.org/10.1155/2020/8872586

Yang, L., Chen, X., Perlaza, S. M., & Zhang, J. (2020). Special issue on artificialintelligence-powered edge computing for Internet of Things. *IEEE Internet of Things Journal, 7*(10), 9224–9226.

Zac, C. H., Rana, M. E., Jayabalan, M., & Hameed, V. A. (2026). Leveraging Cloud Computing for Enhanced Artificial Intelligence and Machine Learning: A Comprehensive Analysis of Service Models, Deployment Strategies, and Provider Offerings. *Lecture Notes in Networks and Systems*, 559–571. https://doi.org/10.1007/978-981-95-0375-9_42

Zangana, H. M., & Zeebaree, S. R. M. (2024). Distributed Systems for Artificial Intelligence in Cloud Computing: A Review of AI-Powered Applications and Services. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 5*(1), 11–30. https://doi.org/10.34010/injiiscom.v5i1.11883

Zebari, S. R. M., & Yaseen, N. O. (2011). Effects of parallel processing implementation on balanced load-division depending on distributed memory systems. *J. of University of Anbar for Pure Science, 5*(3), 50–56.

Zhang, H., Zhang, H., & Ye, C. (2025). A Reinforcement Learning Method for Cyber Range Cloud Resource Scheduling. *2025 6th International Conference on Artificial Intelligence and Computer Engineering (ICAICE),* 877–882. https://doi.org/10.1109/ICAICE68195.2025.11382420

Zheng, Y., & Jiaqing, Y. (2025). Adaptive Auto-Scaling of Distributed Databases on Cloud Platforms via Reinforcement Learning-based Resource Optimization. *2025 International Conference on Modern Sustainable Systems (CMSS),* 971–977. https://doi.org/10.1109/CMSS66566.2025.11182451

Zhou, J., Liu, B., & Gao, J. (2023). A task scheduling algorithm with deadline constraints for distributed clouds in smart cities. *PeerJ Computer Science, 9*, e1346. https://doi.org/10.7717/peerj-cs.1346

Zhou, Z., Feng, J., Liu, L., Wang, J., & Pei, Q. (2024). Heterogeneous Computation and Resource Allocation for Wireless Edge AI. *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2641–2646. https://doi.org/10.1109/GLOBECOM52923.2024.10901095