World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Check for updates

AI-driven resource management strategies for cloud computing systems, services, and applications

Satyanarayan Kanungo *

Principal Data Engineer (Bigdata and Cloud), USA.

World Journal of Advanced Engineering Technology and Sciences, 2024, 11(02), 559–566

Publication history: Received on 04 March 2024; revised on 15 April 2024; accepted on 17 April 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.11.2.0137

Abstract

Cloud computing is a way for businesses and individuals to It has changed and revolutionized the way we access and use resources. However, efficient resource management in cloud computing systems remains a major challenge due to the scalability, heterogeneity, and dynamic nature of these environments. To address these challenges, artificial intelligence (AI) technology has emerged as an effective solution to improve resource management efficiency. This paper provides an overview of AI-based strategies for efficient resource management in cloud computing systems, services, and applications.

This paper first reviews resource management challenges in cloud computing, including scalability, heterogeneity, quality of service requirements, and cost optimization. Below is an overview of the various AI techniques used for resource management. B. Algorithms for machine learning, reinforcement learning, predictive analytics, natural language processing, and genetic algorithms.

Next, this paper considers specific AI-based strategies for efficient resource management. These strategies include automated resource provisioning and scaling, intelligent workload planning and task allocation, predictive maintenance and fault detection, and energy-efficient resource management. We also present case studies and applications of AI-driven resource management in various cloud computing scenarios, including large-scale cloud providers, edge computing, serverless computing, and container environments.

This paper describes evaluation metrics and performance analysis techniques to evaluate the effectiveness of AI-based resource management approaches. It highlights the importance of ethical considerations, transparency, and explainability in AI-powered resource management systems. Additionally, the integration of AI technologies into existing resource management frameworks is discussed, and future directions are identified, including B. real-time resource optimization and coordination.

Keywords: Artificial Intelligence; Resources; Cloud Computing; Efficiency; Cost of Optimization; Quality of Service (QoS)

1. Introduction

1.1. Overview of Cloud Computing Systems, Services, and Applications:

Cloud computing has revolutionized the way computing resources are provisioned, deployed, and utilized. It involves making a shared pool of computer resources, including networks, servers, storage, apps, and services, available for use whenever needed. Cloud computing provides flexibility, scalability, and cost-effectiveness for organizations and

[*] Corresponding author: Satyanarayan Kanungo

individuals, allowing them to focus on their core business objectives without the need for large infrastructure investments. Cloud services can be categorized as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) and offer varying levels of control and management.

1.2. The importance of resource management in cloud computing:

Cloud computing environments require efficient resource management to ensure optimal resource utilization, meet service level agreements (SLAs), and provide a seamless user experience. Resource management is important. Effective resource management includes tasks such as resource provisioning, workload planning, task assignment, error detection, and performance optimization. Poor resource management can lead to underutilization, overprovisioning, increased costs, poor performance, and decreased user satisfaction. Therefore, effective resource management is essential to maximizing the benefits of cloud computing while minimizing operational costs and ensuring high-quality service delivery.

1.3. Artificial intelligence's role in improving resource management efficiency:

Artificial intelligence (AI) technology has emerged as a powerful tool to improve resource management efficiency in cloud computing systems. AI includes various sub-fields such as machine learning, reinforcement learning, predictive analytics, natural language processing, and genetic algorithms. These technologies enable intelligent decision-making, automation, and optimization based on data analysis, pattern recognition, and learning from historical and real-time data. Using AI, cloud computing systems can dynamically allocate resources, optimize workload scheduling, predict resource requirements, detect anomalies and errors, and perform energy-efficient operations. AI-based resource management enables proactive and adaptive strategies that improve the efficiency, scalability, performance, and cost-effectiveness of cloud computing environments.

1.4. Fog Computing Systems

Fog computing is a distributed computing infrastructure where resource-intensive functions reside between the cloud and data sources, reducing the strain on resources (computing power, network bandwidth, and storage capacity). It's structured. Figure 1 shows a fog computing architecture with virtualization enabled. Fog computing architecture has two characteristics: quality of service and energy-aware delivery in virtualized computing environments.

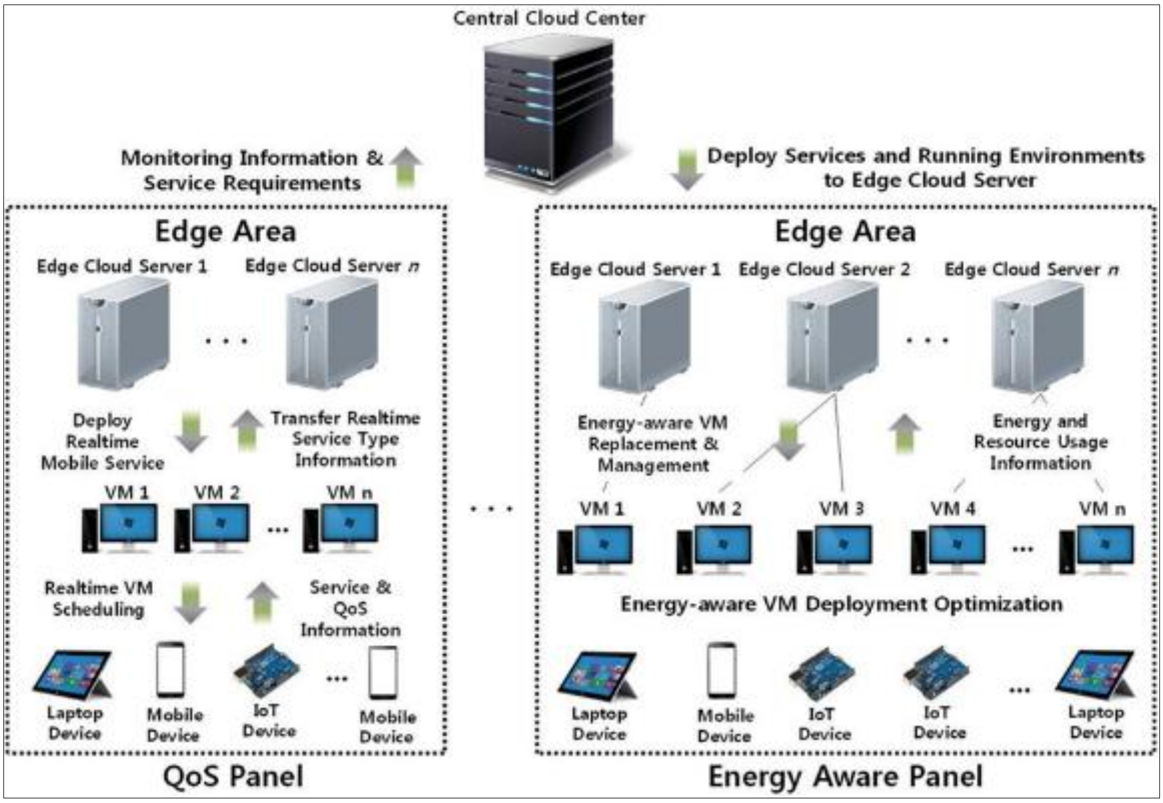

Figure 1 The Fog computing architecture with virtualization enabled.

To ensure the quality of service, the central cloud center interacts with edge cloud servers by obtaining resource monitoring information and service requests. Edge cloud servers provide real-time mobile services on Internet of Things device virtual machines. At the same time, deployed virtual machines send real-time service information, including device type. Internet of Things devices (laptops, smartphones, and sensor devices) interact directly with deployed virtual machines by providing service and quality of service information. Based on the data received, virtual machines are scheduled in real-time.

The resource management strategy of the energy-aware panel in Figure 1 differs from that of the quality of service panel, despite having a comparable architecture. Therefore, the virtual machines send information about their energy and resource usage to the edge cloud server, and the edge cloud server optimizes the use of the virtual machines for energy consumption.
The authors in [12] proposed an intelligent algorithm to offload decisions in fog computing when there are multiple Internet of Things devices in close proximity. The suggested algorithm has two main goals. The suggested algorithm has two main goals. One is device-driven intelligence, and the other is human-driven intelligence for network objectives (energy consumption, latency, network bandwidth, network availability, security/privacy). In this perspective, healthcare research offers an additional fog computing research strategy [13–15]. Fog computing technologies are promising for healthcare applications because they enable the development of predictive techniques for everyday life, where real-time and low latency are important.

1.5. Edge-Cloud Systems

In edge cloud systems, resource functions such as computing power, networking, and storage are distributed throughout the system and closer to the traffic source. Figure 2 shows an edge cloud architecture with mobile devices. The diagram assumes that the mobile device's edge cloud is linked to edge cloud server A. Mobile device (sub-service) tasks from the central cloud server are copied to edge cloud server A. At this point, when the mobile device moves to another location closest to Edge Cloud Server B, the related tasks (sub-services) are scheduled to migrate from Edge Cloud Server A to B.

Mobile phone users should note that their devices are not aware of the central cloud server and edge cloud server processes. Again, when the mobile device moves to another location closest to edge cloud server C, the associated tasks (sub-services) are scheduled to migrate from edge cloud server B to C. This allows mobile and real-time applications to benefit from edge cloud systems with reduced latency.

Regarding intelligent edge cloud systems, the authors [16] proposed an online probabilistic machine learning method that learns from the system's dynamics. This study summarizes wireless communication applications (traffic classification, channel encoding/decoding, channel estimation, planning, and cognitive radio) and proposes an online learning framework for mobile edge computing systems for big data analysis. Masu.

Offloading techniques are widely used because they significantly reduce the computing and communication load on mobile devices. To achieve this objective, the authors in [17] proposed a bidirectional initiative scheme to reduce the decision-making burden. Using a random early detection algorithm, detect network congestion and solve the offloading problem.

For resource management regarding energy consumption in an edge cloud computing environment, Liu et al. addressed resource management concerning energy usage. The proposed framework allows agents to schedule tasks, considering their energy consumption. In contrast to the basic method, Liu et al.'s approach deals with the capacity limitations of edge servers.

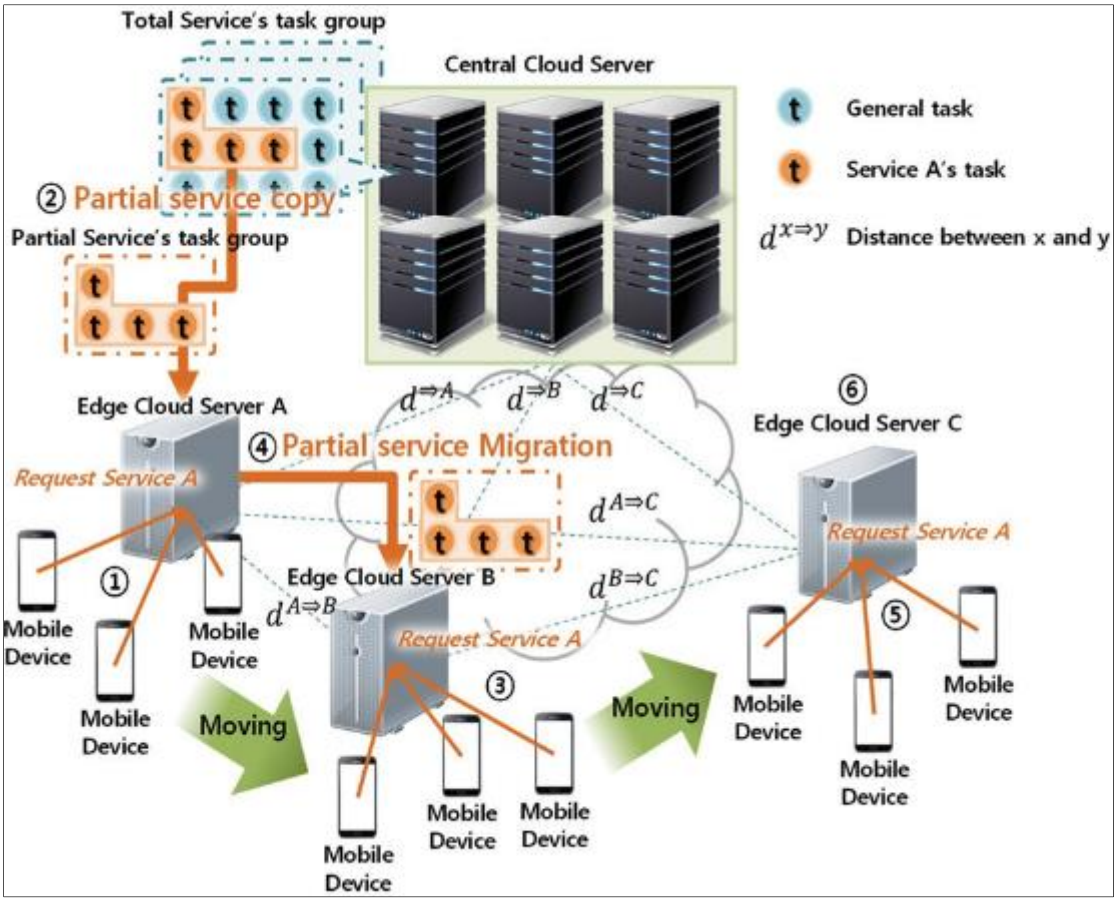

Figure 2 The edge-cloud architecture with mobile devices.

1.6. Intelligent Cloud computing system

Cloud computing is a pay-as-you-go, on-demand model that provides computing resources (CPU, memory, storage, and network) based on virtualization technology. When a user requests a certain amount of computing resources, the cloud data center schedules the deployment according to the request, and the requested virtual machine (or container) is available to him within a minute.

To provide computing resources, cloud data centers aggregate large numbers of physical machines. Therefore, the consolidation of cloud datacenter resources impacts performance and management costs. A well-managed cloud data center reduces energy consumption and carbon emissions.

Artificial intelligence technology can be integrated into cloud computing systems to manage computing resources. The authors in [19] proposed a technique for dynamic resource prediction and allocation in 5G cloud radio access networks. This method uses long-term and short-term memory to predict throughput and uses genetic algorithms to allocate resources.

Chien et al. [20] proposed an intelligent architecture for heterogeneous networks beyond 5G. This architecture's research objective is to use artificial intelligence techniques to enhance network performance in edge cloud computing environments. To maintain quality of service, the authors use packet forwarding techniques and recommend appropriate deep learning techniques for different network topics.

Zhang et al. [21] proposed a multi-algorithm service model to support heterogeneous services and applications. This model is designed to reduce energy consumption and network latency/delay by integrating virtual machines into cloud computing systems. To solve the optimization problem, the authors use a tidal algorithm. The algorithm finds robust results by evaluating the relationship between computational speed and energy cost.

To manage mobile devices in cloud computing environments, we propose an intelligent resource monitoring scheme that predicts future stability based on hidden Markov models. Figure 3 shows the proposed workflow for artificial intelligence applications using hidden Markov chain models. The illustrated workflow is based on an iterative model. Note that other task models can also be applied to the proposed model.

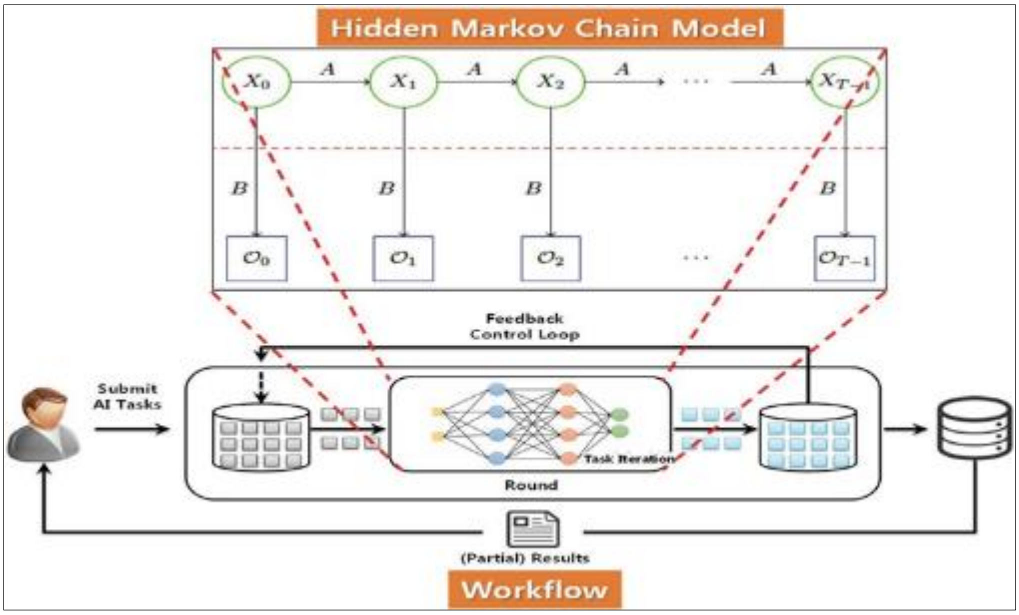

Figure 3 The workflow of artificial intelligence applications with hidden Markov chain model.

When a user submits one or more artificial intelligence tasks, computing resources for the tasks are allocated through the cloud portal system. The allocated resources can be virtual machines, containers, or edge cloud servers, depending on your cloud computing environment. Artificial intelligence tasks are performed by repeating feedback control loops. Once a round is completed, the (partial) results are transferred to the input of the next round.

In the feedback control loop stage, a hidden Markov chain model is applied. Hidden Markov chain models use current and previous mobile device stability information to predict future stability. Specifically, we consider the monitoring information as observable states and calculate the probability of hidden states. Calculating hidden state probabilities allows you to predict your mobile device's future stability. Predicted stability information can be used for cloud integration and cloud resource planning.

An overview and comparison of artificial intelligence-based resource management systems in cloud computing environments can be found in Table 1. Regarding the category of cloud-based systems, our scheme is closely related to intelligent cloud computing systems. Our scheme differs from other studies in terms of features. The proposed intelligent resource management scheme integrates mobile devices into cloud-based systems, "including fog computing and edge clouds." The monitored information. The monitored information The monitored information The monitored information is monitored regularly, and the monitored information is used to predict future stability and mobility based on hidden Markov models. Therefore, our scheme can be used for common cloud applications such as task scheduling, resource consolidation, and computing offloading while optimizing the overall system.

Table 1 Comparison and summary of resource management schemes based on artificial intelligence

Category	Study	Characteristics	Technique/consideration	Application
	13	Energy and latency reduction	Machine learning, task offloading	Body sensor network, health monitoring
	14	Achieve overall system performance	Geo-distributed system between sensor nodes and cloud	Heathcare, smart home
	15	Sensitive data protection, delay reduction	Patient driven healthcare architecture	Healthcare (individual clustered)

Edge-cloud	16	Decoupled of tasks between time slots and edge devices	Machine learning for wireless communication	Mobile edge computing, big data analytics
	17	Avoidance of network congestion	Computation offloading, wired/wireless communication	Task scheduling in edge- cloud systems
	18	Improvement of the energy management performance, reduction of the execution time	Energy-aware scheduling scheme with deep reinforcement learning	Smart cities (smart building, smart power grid, multi-energy networks)
Intelligent cloud computing	19	Improvement of chip assembly and production efficiency	Cognitive manufacturing, intelligent manufacturing	Robot-factory
	20	Implementation of intelligent system architectures and network function	Heterogeneity of beyond 5G	Resource allocation, integrated packet forwarding
	21	Optimization of energy consumption and delay	Workload weights and the computation capacities	Artificial intelligence applications

Ours predict future stability and mobility Hidden Markov model Mobile and artificial intelligence application.

2. Conclusion

In summary, the rapid growth of cloud computing systems, services, and applications requires the development of efficient resource management strategies to meet the increasing demands in this dynamic environment. In this research article, we explored the potential of AI-driven approaches to address challenges related to resource allocation, optimization, and scalability in cloud computing.

By harnessing the power of artificial intelligence, cloud computing systems can benefit from intelligent decision-making, predictive analytics, and automated resource provisioning. AI algorithms such as machine learning and deep learning analyze large amounts of data, identify patterns, and make informed predictions to optimize resource utilization, improve performance, and reduce costs. Masu.

Integrating AI-driven resource management strategies into cloud computing offers many benefits. This enables proactive resource allocation, where resources are dynamically allocated based on workload patterns, user needs, and performance requirements. This not only ensures optimal utilization, but also increases system reliability and responsiveness.

AI algorithms also enable predictive scaling, allowing cloud providers to predict future resource needs and scale accordingly. By leveraging historical data, workload forecasting, and predictive analytics, cloud systems can proactively adjust resource allocation based on fluctuating demand, avoiding potential bottlenecks and over-provisioning.

Furthermore, AI-powered resource management strategies contribute to improving security and fault tolerance in cloud computing. Machine learning algorithms can detect anomalies, identify potential security threats, and take proactive steps to reduce risk. Automated failure detection and recovery mechanisms can be used to ensure system resiliency and minimize downtime.

However, it is important to be aware of the challenges and considerations associated with AI-driven resource management in cloud computing. Ethical concerns, privacy, algorithmic bias, and interpretability are important factors that need to be carefully considered to ensure fairness, transparency, and accountability.

In summary, AI-driven resource management strategies have the potential to revolutionize cloud computing systems, services, and applications. These strategies optimize resource utilization, improve performance, increase scalability, and strengthen security through intelligent decision-making, predictive analytics, and proactive resource allocation. As cloud computing continues to evolve and expand, enterprises must adopt an AI-driven resource management approach to remain competitive, provide efficient services, and meet the ever-increasing demands of the digital age. It becomes important.

References

[1] Yao, M., Sohul, M., Marojevic, V., & Reed, J. H. (2019). Artificial intelligence defined 5G radio access networks. IEEE Communications Magazine, 57(3), 14–20.

[2] ur Rehman, M. H., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P. P., & Perera, C. (2019). The role of big data analytics in industrial Internet of Things. Future Generation Computer Systems, 99, 247–259.

[3] Zhang, Y., Ma, X., Zhang, J., Hossain, M. S., Muhammad, G., & Amin, S. U. (2019). Edge intelligence in the cognitive Internet of Things: Improving sensitivity and interactivity. IEEE Network, 33(3), 58–64.

[4] Sodhro, A. H., Pirbhulal, S., & de Albuquerque, V. H. C. (2019). Artificial intelligence-driven mechanism for edge computing-based industrial applications. IEEE Transactions on Industrial Informatics, 15(7), 4235–4243.

[5] Dai, Y., Xu, D., Maharjan, S., Qiao, G., & Zhang, Y. (2019). Artificial intelligence empowered edge computing and caching for Internet of vehicles. IEEE Wireless Communications, 26(3), 12–18.

[6] Donida Labati, R., Genovese, A., Piuri, V., Scotti, F., & Vishwakarma, S. (2020). Computational intelligence in cloud computing. In Recent Advances in Intelligent Engineering (pp. 111–127). Cham: Springer International Publishing.

[7] Satyanarayanan, M., & Davies, N. (2019). Augmenting cognition through edge computing. Computer, 52(7), 37–46.

[8] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25.

[9] Gacanin, H., & Wagner, M. (2019). Artificial intelligence paradigm for customer experience management in next-generation networks: Challenges and perspectives. IEEE Network, 33(2), 188–194.

[10] Chien, W. C., Lai, C. F., & Chao, H. C. (2019). Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence. IEEE Transactions on Industrial Informatics, 15(7), 4306–4314.

[11] Li, Z., Liu, L., & Kong, D. (2019). Virtual machine failure prediction method based on AdaBoost-Hidden Markov model. In Proceedings of 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 700–703). Changsha, China.

[12] Mutlag, A. A., Abd Ghani, M. K., Arunkumar, N., Mohammed, M. A., & Mohd, O. (2019). Enabling technologies for fog computing in healthcare IoT systems. Future Generation Computer Systems, 90, 62–78.

[13] La, Q. D., Ngo, M. V., Dinh, T. Q., Quek, T. Q. S., & Shin, H. (2019). Enabling intelligence in fog computing to achieve energy and latency reduction. Digital Communications and Networks, 5(1), 3–9.

[14] Rahmani, A. M., Gia, T. N., Negash, B., Anzanpour, A., Azimi, I., Jiang, M., & Liljeberg, P. (2018). Exploiting smart e-health gateways at the edge of healthcare Internet-of-Things: A fog computing approach. Future Generation Computer Systems, 78, 641–658.

[15] Kumari, A., Tanwar, S., Tyagi, S., & Kumar, N. (2018). Fog computing for Healthcare 4.0 environment: Opportunities and challenges. Computers & Electrical Engineering, 72, 1–13.

[16] Cui, Q., Gong, Z., Ni, W., Hou, Y., Chen, X., Tao, X., & Zhang, P. (2019). Stochastic online learning for mobile edge computing: Learning from changes. IEEE Communications Magazine, 57(3), 63–69.

[17] Yin, Z., Chen, H., & Hu, F. (2019). An advanced decision model enabling two-way initiative offloading in edge computing. Future Generation Computer Systems, 90, 39–48.

[18] Liu, Y., Yang, C., Jiang, L., Xie, S., & Zhang, Y. (2019). Intelligent edge computing for IoT-based energy management in smart cities. IEEE Network, 33(2), 111–117.

[19] Hu, L., Miao, Y., Wu, G., Hassan, M. M., & Humar, I. (2019). iRobot-Factory: An intelligent robot factory based on cognitive manufacturing and edge computing. Future Generation Computer Systems, 90, 569–577

[20] Chien, W. C., Cho, H. H., Lai, C. F., Tseng, F. H., Chao, H. C., Hassan, M. M., & Alelaiwi, A. (2019). Intelligent architecture for mobile HetNet in B5G. IEEE Network, 33(3), 34–41.

[21] Zhang, W., Zhang, Z., Zeadally, S., Chao, H. C., & Leung, V. C. M. (2019). MASM: A multiple-algorithm service model for energy-delay optimization in edge artificial intelligence. IEEE Transactions on Industrial Informatics, 15(7), 4216–4224.

[22] Lim, J. B., Lee, D. W., Chung, K. S., & Yu, H. C. (2019, January 1). Intelligent Resource Management Schemes for Systems, Services, and Applications of Cloud Computing Based on Artificial Intelligence. https://doi.org/10.3745/jips.04.0139