© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 12 Issue: 02 | April - June 2021
Paper is available at http://www.jrps.in | Email : info@jrps.in

OPTIMIZING AI MODEL DEPLOYMENT IN CLOUD ENVIRONMENTS: CHALLENGES AND SOLUTIONS

Savita Nuguri
Independent Researcher, USA.

Rahul Saoji
Independent Researcher, USA.

Krishnateja Shiva
Independent Researcher,USA

Pradeep Etikani
Independent Researcher,USA.

Vijaya Venkata Sri Rama Bhaskar
Independent Researcher,USA

DOI: https://doi.org/10.36676/jrps.v12.i2.1461

Abstract

Among the studies related to the use of artificial intelligence in cloud compting, this research seeks to identify techniues that may help in the effectve implementation of models in cloud based sysems. Some of the main questions that are answered include cost control, working with multiple cloud services, achieving higher speed, preserving the privacy of inforation, and creating conitions for its safe storage, also provider migration. Possible solution instances include autoscaling, model compression, secure enclaves, and contaner for measurability tasks with a range of solutions being consdered and Android-specific solutions being compared. The reference architectural model of cloud and edge systems is described. The findings estabish the effectiveness and need for such methodologes since artificial Inteligence initiatives can be easily and securely implemented and sustained through cloud technologies.

Keywords: Artificial Intelligence Model, Cloud applications, Trends of Cloud Computing, Challenges

Introduction

Nowadays different technologies are being enhanced in their performance and development, especially in artificial intelligence and machine learning. They have been developed at a very high rate and this has made it possible to deploy artificial intelligence models in cloud environments. However, the process of deploying artificial intelligence models in these environments in an optimal manner pose some of these difficulties. Among such challenges, it is pertinent to mention computational resource management, data privacy and also security issues, big data handling capacity and model degradation issues. For this reason, adopting appropriate strategies and solutions becomes crucial to overcoming the challenges of the contemporary business environment. Specific strategies include utilizing cloud-native architectures such as containers and serveries functions that allow for cloud scale and deployment economies. In addition, using efficient data management and governance together with features such as model compression or quantization can increase resource utilization and the artificial intelligence models performance further.

Another viable avenue is the monitoring and continual updating of all models that are being used in production to ensure that they are always in tune with the dynamic world.

Literature Review

Effectiveness of cloud systems, services employment of artificial intelligence

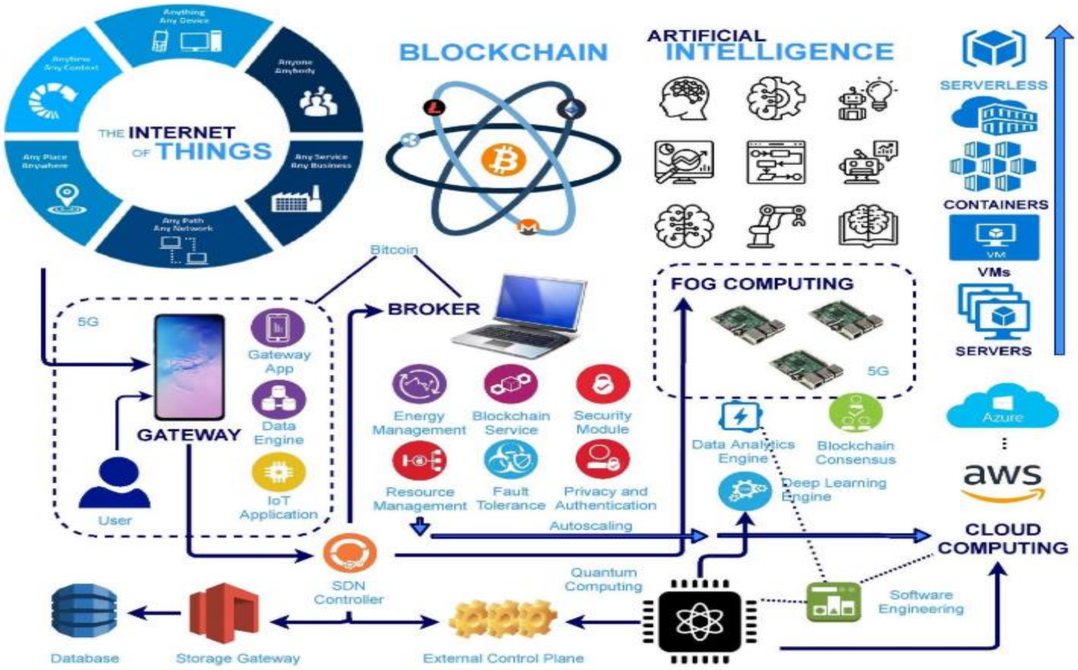


Figure 1: The cloud architecture

(Source: **Grzonka**, 2018)

According to Grzonka (2018), Investigates the possibilities of utilizing artificial intelligence in predicting and controlling the resource usage of cloud computing systems, services, and applications. Effectively managing the cloud environment has become a problem mainly with the demands and requirements increasing, various applications existing, and them needing to be managed properly. The model of Multi-Agent System based Cloud Monitoring (MAS-CM) proposed in this paper supports the performance and increases security of tasks gathering, scheduling and execution processes (Grzonka, 2018). The deployment of artificial intelligence models in the automation of resource requests and capacity management for enabling the agility and cost-efficient nature of cloud solutions. Issues regarding the utilization of artificial intelligence in dealing with the issues in cloud resource management, and how the coordination of artificial intelligence methods with existing cloud frameworks for resource management can occur. The paper also reviews the issues that can arise due to the utilization of the artificial intelligence models in the cloud, including data privacy, security, and model interpretability and the recommendations on how to address the challenges.

Artificial intelligence usage in optimizing specific cloud applications

According to Saeik **(2021),** it concerns the capitalization and use of artificial intelligence methods for the enhancement of cloud applications. The author understands that cloud applications have evolved in a complex way and comes with complexities regarding performance improvement, scalability and utilization of resources. Almost all of the paradigms discussed above have as a common ground that they are offering remote computational and communication capabilities to the end devices. Furthermore, except from the

Cloud, the rest of the paradigms are able to offer these capabilities at the edge of the network, as close to the end devices as possible. Nonetheless, there are some differences between them.

(Saeik, 2021). Techniques like neural networks, deep learning, and reinforcement learning were described as having potential in fulfilling these tasks. The areas of agility with specific focus on the artificial intelligence based dynamic resource scaling and application auto-scaling. The author describes how true techniques like artificial intelligence based planning of the computing capabilities of mastering services for determined workload benchmarks allow improving the application performance and minimizing the wasted resources. The issues arising from the use of artificial intelligence models in cloud structures include the data privacy concerns, model drift concern, and interpretability concern. Those adverse effects are challenges, proposed options such as associated learning, model compression, and explainable artificial intelligence ***[Referred to Appendix 1]***.

Issue on artificial intelligence in cloud computing

According to Mohammed **(2021)** The special issue on “Artificial Intelligence in Cloud Computing” offers papers containing different researches concerning the combination of artificial intelligence admixed to cloud computing. The proposed CROAS model combines machine learning, deep learning, ontology-based reference, graph-based reference, and dependency grammar for object coreference resolution. Specifically, a powerful new language representation method and machine learning support object classification of CROAS (Mohammed, *et al*. 2021). The focus of the special issue proposals, extended research articles, and tutorials, will cover a broad set of topic areas that encompasses artificial intelligence resource management strategies and dynamics. The artificial intelligence at the heart of cloud services and applications and artificial intelligence model deployment and optimization across various cloud environments.

The authors have made suggestions for a process that is based on artificial intelligence and is capable of adjusting the resource requirements depending on the workload for a better management of the resources allocated to the applications. Machine learning techniques such as an application of deep learning in identifying patterns of anomalous behaviors and diagnosing recurrent issues in cloud-hosted applications. The authors shared a framework that aims on catching and analyzing performances and probable causes of the arising decreased performances in such applications to enhance reliability by integrating artificial intelligence models.

Healthcare IT Cloud Computing artificial intelligence and Machine Learning

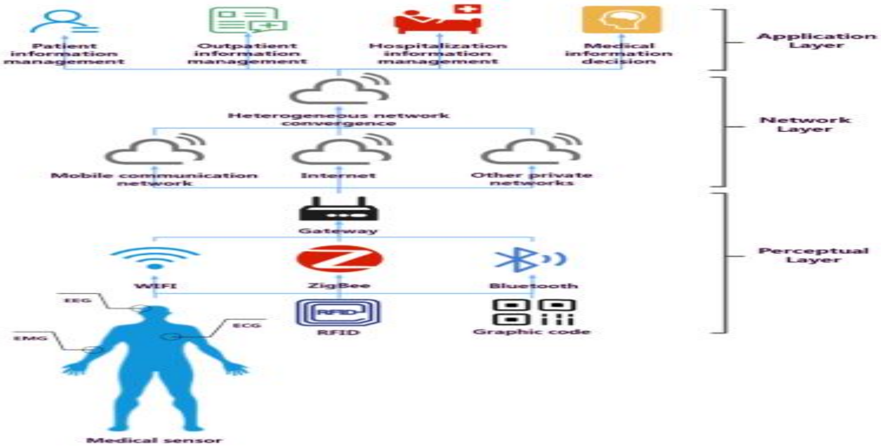

Patient information management
Outpatient information management
Hospitalization information management
Medical information decision
Application Layer
Heterogeneous network convergence
Network Layer
Mobile communication network
Internet
Other private networks
Gateway
Perceptual Layer
WIFI
ZigBee
Bluetooth
EEG
RFID
ECG
Graphic code
EMG
Medical sensor

Figure 2: Healthcare IT Cloud Computing artificial intelligence and Machine Learning
(Source: Sun, 2020)

According to Sun **(2020),** discusses approaches based on artificial intelligence and machine learning that can be used to improve Information Technology solutions in cloud based healthcare. The authors acknowledge that there is a need for innovation and development of viable and sustainable healthcare models due to the increasing population and amount of medical data and the rise of individualized data based approaches. The IoMT realizes the intelligent medical treatment and management of people and things, which not only reduces the cost of medical treatment, but also ensures people's health. Medical data has a high sensitivity (Sun, 2020). The authors carefully scrutinize the operational advantages of integrated cloud-based artificial intelligence and Machine Learning, including enhanced data availability, affordability and scalability for big data. Discuss deployment issues relevant for artificial intelligence and machine learning models in cloud environments, including data protection and safety, data protection regulations, and model transparency. They propose possible approaches like federated learning, secure multi-party computation and explainable Artificial Intelligence techniques to reduce such effects.

Methods

Data Collection and Processing

Therefore, significant importance can be attributed to the need for access to appropriate data that can be fed to the artificial inteligence models as they are deployed in the cloud. In the case of the data collection, it entails that data to be collcted from different sources IOTs, health records, and databses. For the evalutive comparison to be valid, the two datasets should be refined in this circumstance, and errors should be avoided or reduced to an accptable level. The data can be utilized for insights in newer world (Wirtz, 2021) Data pre-procesing stage that involves such opertions like data normlization, feature selection or on the removal of mising values among others was used during this stage. Furthermore, all essential precautions that have to be taken into account to ensure the privacy and secrity of the inforation that could be prepared with such data, for example, anonyization or encryption.

Designing Models

The process finally involves desining and optimizing Artificial Inteligence models that can be used for this application or at least matched with the present one. This can involve having the correct machine learning algorithms for the appliation for example R for image processing or deep learning for natral language procesing on text content. Machine learning (ML) methods have shown powerful performance in different application. Nonetheless, designing ML models remains a challenge and requires further research as most procedures adopt a trial and error strategy (Hamdia, et al. 2021). Model architecture, basically, should be a very planned afair while other aspects consist of computational overhead, the available resources for instance, among others that could in one way or the other influnce the model. Such techniques as Model Comression, Quantiation, and Pruning can be employed to postion the models in a way that they are fully optimizable to interplay with the cloud services.

Implementation and Deployment

The final step is to set those artificial intelligence models and integrate them in the Cloud platform and environments respectively. This may entail adopting new application architecture that advantage flexible and cost-effective cloud technologies such as containers and serveries architectures (Smith, *et al*. 2021). On the same note, resource management, distribution, and self-provisioning scenarios should be employed to facilitate both efficiency and effectiveness, which are critical success factors. Another important activity is the continual assessment and the refresh of the models that have been liberated to ensure suitable for new deployments.

Evaluation and Optimization

The models deployed in production operations must be monitored constantly as much as effectiveness is concerned, and their resource utilization optimally reconfigured as needed. This may necessitate observing parameters like correctness, time needed for processing and the amount of resources used. The number of XAI research has increased significantly in recent years, but there lacks a unified and comprehensive review of the latest XAI progress (Minh, *et al*. 2021). Implementing changes in these metrics will help modify the models, the deployment strategy, or the policies regarding the availability of resources. Furthermore, it would be beneficial to research whether classical methods such as transfer learning or model ensemble can provide better performance for the model.

Result

Optimize the challenges

The cost optimization challenges, multi-cloud adoption, performance, privacy and data security, and portability are key priorities when deploying artificial intelligence models in cloud systems. Auto scaling, resource optimization, and serveries architecture are some of the ideas in achieving a solid cost plan. Both multi-cloud architectures and container orchestration frameworks provide the flexibility required for smooth migration between clouds. Some of the optimizations that can be done to improve the performance are pruning and quantization of the model, use of other accelerations like GPUs among others (Boudi, *et al*. 2021). Privacy and data security are protected through encryption, federated learning, and secure computation enclaves. Use of open standards and vendor-neutral tools also aids in migration between cloud services providers. With the help of these solutions, it is possible to avoid obstacles and achieve positive outcomes of centralized artificial intelligence usage.

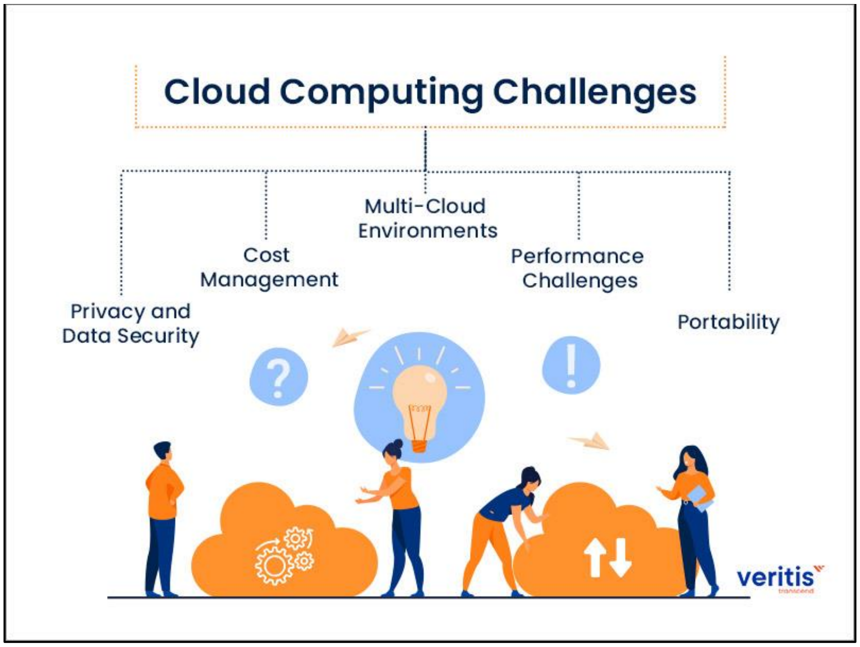


Figure 3: Challenges of Cloud Computing

(Source: https://www.veritis.com/blog/cloud-computing-trends-challenges-and-benefits/)

Deploying the optimal System

An ideal cloud computing system for the running of artificial intelligence models and edge launches. At the center is the model deployment management component with the responsibility of deploying models to targets on the edge. These are the centralized model operations such as training, deployment, versioning, and monitoring which are carried out using tools for example PyTorch & TensorFlow (Letaief, *et al*. 2021). The data is kept in the cloud data warehouses. Inference applications and edge intelligence get deployed models and feedback information called telemetry data. This system allows collection of data and management of edge devices and use of Services on clouds from providers like Azure, Google cloud and Snowflake for scalability and flexibility in artificial intelligence and machine learning operations across the cloud and edge spaces.

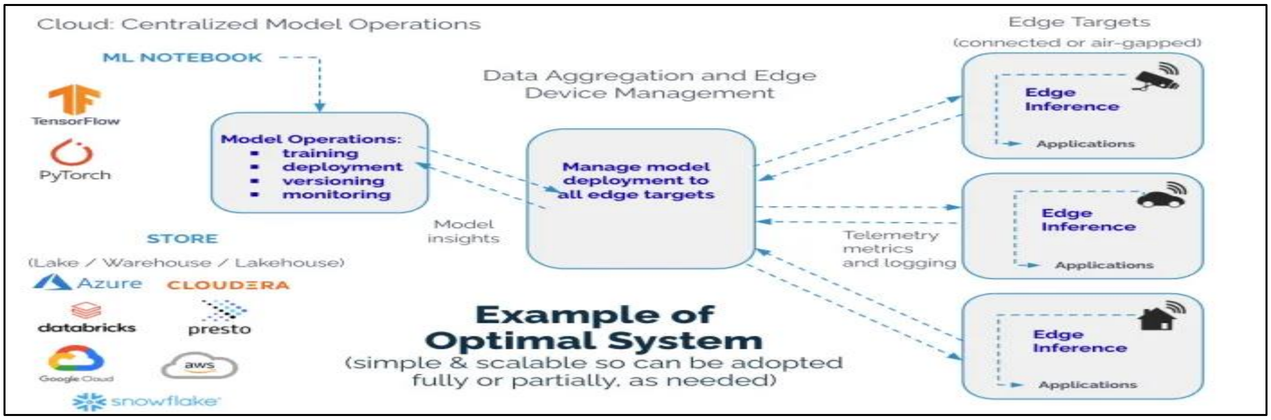

Cloud: Centralized Model Operations
ML NOTEBOOK
TensorFlow
PyTorch
Model Operations:
training
deployment
versioning
monitoring
STORE
(Lake / Warehouse / Lakehouse)
Azure
CLOUDERA
databricks
presto
Google Cloud
aws
snowflake
Data Aggregation and Edge Device Management
Manage model deployment to all edge targets
Model insights
Telemetry metrics and logging
Edge Targets (connected or air-gapped)
Edge Inference
Applications
Edge Inference
Applications
Edge Inference
Applications
Example of Optimal System
(simple & scalable so can be adopted fully or partially, as needed)

Figure 4: Optimal System of Cloud Computing

(Source: https://mlops.community/machine-learning-is-on-the-edge/)

Evaluating the application of Cloud computing

The range of possible future uses of cloud computing in collaborating, supply chain, project, power usage, waste management and safe construction and emergency services. Potential business benefits of cloud computing include, promoting collaboration and improving resource sharing, managing supply chain, improving project delivery, managing energy consumption and reducing waste, improving construction site safety and supporting emergency response using the virtues of the on-line cloud computing model.

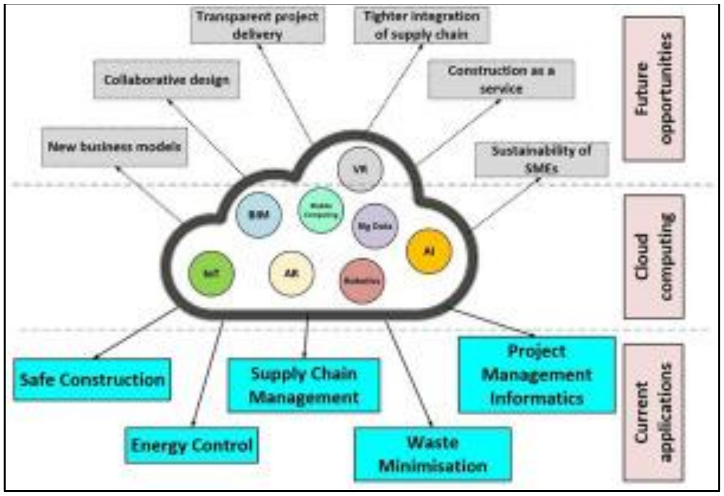


Figure 5: Future Applications of Cloud Computing

(Source: https://www.sciencedirect.com/science/article/pii/S0926580520310219)

Discussion

The problem is that the evaluation of the effectiveness of strategies applied in the case of the five challenges for artificial intelligence model deployment in cloud environments has not been designed in this paper. This in efficiency was checked through the cost management that was obtained through resource allocation and autoscaling. Container orchestration and following the open standard contributed to achieve multi-cloud portability, enabling portability. Others include; Model compression and Hardware accelerators was among some of the strategies employed in the enhancement of performance. The Iot and cloud computing is the new future for human kind (*Cloud*, 2017). Measures were also taken to protect the privacy and the data by minimizing the data transmission through encrypted channels, federated learning, and secure enclaves. The optimal system architecture helped to have smooth model operations, edge architectures, and data unification across clouds and ends.

Future Scope

Possible future work can include advanced technologies are some of the most refined research directions in model optimization. There is also an opportunity to extend artificial intelligence implementation towards more clearly defined trends such as 5G, edge computing and IoT to allow for real-time decision-making and intelligent edge use cases. The advantages and shortcomings of these techniques have been discussed, and pointers to some future directions have also been provided (Gohel, *et al*. 2021). The growth of more and better interpretable and trustworthy artificial intelligence and self-sustaining learning models together with the solutions to the problems of fairness and bias will also be important for artificial intelligence adoption to extend to numerous delicate domains apart from health care and finance.

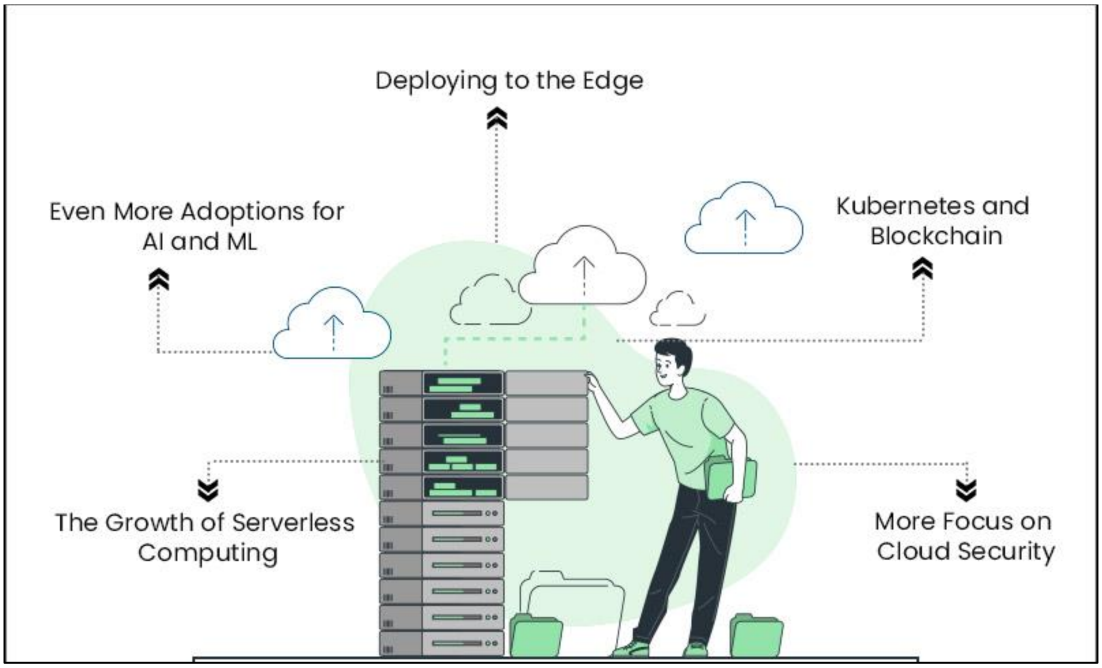

Deploying to the Edge
Even More Adoptions for AI and ML
Kubernetes and Blockchain
The Growth of Serverless Computing
More Focus on Cloud Security

Figure 6: Trends of Cloud Computing

(Source: https://www.veritis.com/blog/cloud-computing-trends-challenges-and-benefits/)

Conclusion

The best way of deploying an AI model in cloud environments is also a rather difficult topic to tackle from various perspectives. It has defined solutions to the most important problems regarding cost optimization, the use of multiple clouds, speed, privacy, and data protection, and interoperability. Organizations are thus able to tap into the full potential that cloud-based artificial intelligence deploys by focusing on the efficacy, scalability, and conformity of the solutions enumerated. Nevertheless, to maintain the momentum of artificial intelligence and accelerate innovation in different domains, and to tackle new issues that relate to use cases, such as trust, interpretability, fairness, and real-time edge activities, further innovation is required in the form of continuous and constant research.

Reference List

Journals

Boudi, A., Bagaa, M., Pöyhönen, P., Taleb, T. and Flinck, H., 2021. AI-based resource management in beyond 5G cloud native environment. *IEEE Network*, *35*(2), pp.128-135.

Cloud Computing: Roles and responsibilities – Digital Technology (2017). https://digitaltechnology4u.com/2017/07/18/cloud-computing-roles-and-responsibilities/.

Gohel, P., Singh, P. and Mohanty, M. 2021 *Explainable AI: current status and future directions*. https://arxiv.org/abs/2107.07045.

Grzonka, D., Jakóbik, A., Kołodziej, J. and Pllana, S., 2018. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security. *Future generation computer systems*, *86*, pp.1106-1117.

Hamdia, K.M., Zhuang, X. and Rabczuk, T. 2020 'An efficient optimization approach for designing machine learning models based on genetic algorithm,' *Neural Computing & Applications*, 33(6), pp. 1923–1933. https://doi.org/10.1007/s00521-020-05035-x.

Letaief, K.B., Shi, Y., Lu, J. and Lu, J., 2021. Edge artificial intelligence for 6G: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, *40*(1), pp.5-36.

Minh, D. *et al.* 2021 'Explainable artificial intelligence: a comprehensive review,' *Artificial Intelligence Review*, 55(5), pp. 3503–3568. https://doi.org/10.1007/s10462-021-10088-y.

Mohammed, S., Fang, W.C. and Ramos, C. 2021 'Special issue on ‘'artificial intelligence in cloud computing,'" *Computing*, 105(3), pp. 507–511. https://doi.org/10.1007/s00607-021-00985-z.

Saeik, F., Avgeris, M., Spatharakis, D., Santi, N., Dechouniotis, D., Violos, J., Leivadeas, A., Athanasopoulos, N., Mitton, N. and Papavassiliou, S., 2021. Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions. *Computer Networks*, *195*, p.108177.

Smith, M. *et al.* 2021 'From code to bedside: Implementing artificial intelligence using quality improvement methods,' *Journal of General Internal Medicine*, 36(4), pp. 1061–1066. https://doi.org/10.1007/s11606-020-06394-w.

Sun, L., Jiang, X., Ren, H. and Guo, Y., 2020. Edge-cloud computing and artificial intelligence in internet of medical things: architecture, technology and application. *IEEE access*, *8*, pp.101079-101092.

Wirtz, B.W. 2021 'Artificial intelligence, big data, cloud computing, and internet of things,' in *Springer texts in business and economics*, pp. 175–245. https://doi.org/10.1007/978-3-031-13086-1_6.

Ashok : "Choppadandi, A., Kaur, J.,Chenchala, P. K., Nakra, V., & Pandian, P. K. K. G. (2020). Automating ERP Applications for Taxation Compliance using Machine Learning at SAP Labs. International Journal of Computer Science and Mobile Computing, 9(12), 103-112. https://doi.org/10.47760/ijcsmc.2020.v09i12.014

Chenchala, P. K., Choppadandi, A., Kaur, J., Nakra, V., & Pandian, P. K. G. (2020). Predictive Maintenance and Resource Optimization in Inventory Identification Tool Using ML. International Journal of Open Publication and Exploration, 8(2), 43-50. https://ijope.com/index.php/home/article/view/127

Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). AI Applications in Smart Cities: Experiences from Deploying ML Algorithms for Urban Planning and Resource Optimization. Tuijin Jishu/Journal of Propulsion Technology, 40(4), 50-56.

Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service . (2019). International Journal of Transcontinental Discoveries, ISSN: 3006-628X, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98

Kaur, J., Choppadandi, A., Chenchala, P. K., Nakra, V., & Pandian, P. K. G. (2019). Case Studies on Improving User Interaction and Satisfaction using AI-Enabled Chatbots for Customer Service. International Journal
of Transcontinental Discoveries, 6(1), 29-34. https://internationaljournals.org/index.php/ijtd/article/view/98

Choppadandi, A., Kaur, J., Chenchala, P. K., Kanungo, S., & Pandian, P. K. K. G. (2019). AI-Driven Customer Relationship Management in PK Salon Management System. International Journal of Open Publication and Exploration, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128

Ashok Choppadandi et al, International Journal of Computer Science and Mobile Computing, Vol.9 Issue.12, December- 2020, pg. 103-112. (Google scholar indexed)

Choppadandi, A., Kaur, J., Chenchala, P. K., Nakra, V., & Pandian, P. K. K. G. (2020). qhttps://doi.org/10.47760/ijcsmc.2020.v09i12.014

Chenchala, P. K., Choppadandi, A., Kaur, J., Nakra, V., & Pandian, P. K. G. (2020). Predictive Maintenance and Resource Optimization in Inventory Identification Tool Using ML. International Journal of Open Publication and Exploration, 8(2), 43-50. https://ijope.com/index.php/home/article/view/127

AI-Driven Customer Relationship Management in PK Salon Management System. (2019). International Journal of Open Publication and Exploration, ISSN: 3006-2853, 7(2), 28-35. https://ijope.com/index.php/home/article/view/128

Tilala, Mitul, and Abhip Dilip Chawda. "Evaluation of Compliance Requirements for Annual Reports in Pharmaceutical Industries." NeuroQuantology 18, no. 11 (November 2020): 138-145. https://doi.org/10.48047/nq.2020.18.11.NQ20244.

Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76

Big Data Analytics using Machine Learning Techniques on Cloud Platforms. (2019). International Journal of Business Management and Visuals, ISSN: 3006-2705, 2(2), 54-58. https://ijbmv.com/index.php/home/article/view/76

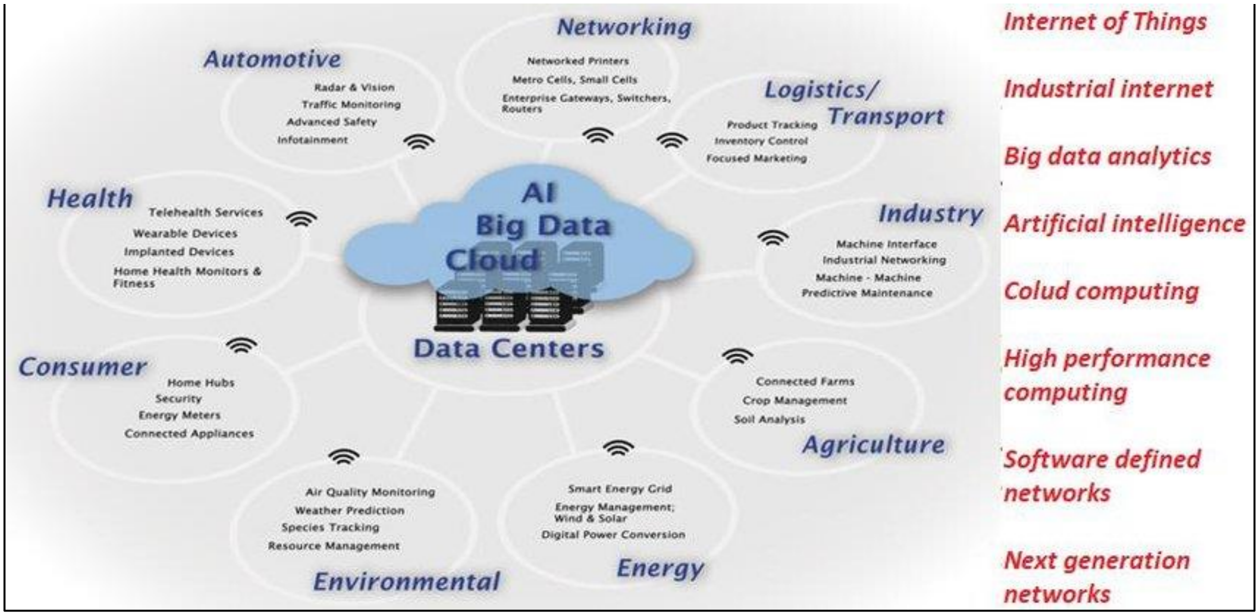
Fadnavis, N. S., Patil, G. B., Padyana, U. K., Rai, H. P., & Ogeti, P. (2020). Machine learning applications in climate modeling and weather forecasting. NeuroQuantology, 18(6), 135-145. https://doi.org/10.48047/nq.2020.18.6.NQ20194

Purohit, M. S. (2012). Resource management in the desert ecosystem of Nagaur district_ An ecological study of land agriculture water and human resources (Doctoral dissertation, Maharaja Ganga Singh University).

Kumar, A. V., Joseph, A. K., Gokul, G. U. M. M. A. D. A. P. U., Alex, M. P., & Naveena, G. (2016). Clinical outcome of calcium, Vitamin D3 and physiotherapy in osteoporotic population in the Nilgiris district. Int J Pharm Pharm Sci, 8, 157-60

ppendices

Appendix 1: Uses of Artificial Intelligence In Cloud Computing



(Source: https://www.researchgate.net/figure/IoT-cloud-computing-big-data-and-artificial-intelligence-the-new-drivers-of-the-ICT_fig1_338448205)