World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

ML-driven resource management in cloud computing

Tanvir Mahmud *

Department of EEE, Daffodil International University, Daffodil Smart City (DSC), Dhaka-1216, Dhaka, Bangladesh.

World Journal of Advanced Research and Reviews, 2022, 16(03), 1230-1238

Publication history: Received on 11 November 2022; revised on 20 December 2022; accepted on 23 December 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.16.3.1398

Abstract

This paper explores the challenges associated with cloud resource management, the application of ML techniques to address these challenges, and their associated benefits and limitations. Key ML applications in cloud computing include workload prediction, energy-efficient VM consolidation, QoS-aware resource provisioning, and network-aware VM placement. The study also identifies research gaps and proposes future directions for enhancing ML-driven resource management in cloud environments, with a focus on deep learning, reinforcement learning, and ensemble methods. By leveraging ML, cloud computing systems can achieve improved scalability, cost-effectiveness, and performance, paving the way for next-generation intelligent cloud infrastructure.

Keywords: Cloud Computing; Machine Learning; Resource Management; Workload Prediction; VM Consolidation; Energy Efficiency; Deep Learning; Reinforcement Learning; QoS Optimization.

1. Introduction

In July 2020, in demographic research conducted by Statista, approximately 4.57 billion users were active on the Internet, encompassing 59% of the global population [1]. This huge number of users shows the demanding need for resilient, secure and easily configurable web applications. Before the cloud computing era, from the enterprise perspective, the cost of the maintenance of big data centers and bootstrap ping a company was too high. Cloud Computing is the provision of virtual resources via the Internet (e.g., servers, apps, etc.), from central systems located away from the end users, which serves them by automating processes, providing convenience, flexibility of connection [2] as well as a nice pay-as-you-go payment plan to the company [2, 3].

2. Cloud computing

Cloud computing provides resources over the Internet, such as memory, CPU, bandwidth, disc, and applications/services. The National Institute of Standards and Technology (NIST) [4] states that "Cloud computing is a model for providing on-demand network access to a common pool of configurable computing resources (e.g., networks, servers, storage, software, and services) that can be quickly provisioned and released with minimal management effort or service provider involvement. There are five core features, three service models, and four deployment options in this cloud model''. Based on the literature, two more characteristics have been included.

This computing model uses a client–server architecture to centralize application deployment and computation offloading. Cloud computing is cost-effective in application delivery and maintenance on both the client and server sides and flexible in resource provisioning and detaching services from related technologies. Cloud computing and its supporting technology have been investigated for years. Many advanced computing systems have been released to the market, including Alibaba Cloud, Microsoft Azure, Adobe Creative Cloud, ServerSpace, Amazon Web Services (AWS), and Oracle Cloud [5].

* Corresponding author: Tanvir Mahmud

3. Cloud computing service models

• Software as a Service (SaaS) [6]: Using this service model, a client can access the service provider Cloud hosted applications. Web portals are used to access applications. Since providers have access to the applications, this model has made production and testing easier for them.

• Platform as a Service (PaaS) [7]: In this service model, the service provider provides basic requirements including network, servers, and operating system to enable the client to build acquired applications and manage their configuration settings.

• Infrastructure as a Service (IaaS) [8]: The user has created all necessary applications and only requires a simple infrastructure. Vendors may include processors, networks, and storage as facilities with customer provisions in such cases.

4. Deployment Models for Cloud Computing

Public Cloud [9]: The public cloud is the most widely used cloud computing model, where service providers offer cloud-based resources over the Internet following predefined policies, regulations, and business models. Due to the extensive shared resource pool, public cloud providers can offer customers a variety of options while ensuring Quality of Service (QoS).

Private Cloud [10]: A private cloud is designed to serve a single organization or institution, providing many of the benefits of a public cloud while offering enhanced security through corporate firewalls. However, the cost of setting up and maintaining a private cloud is significantly high, as the organization managing it is responsible for all aspects of the system.

Community Cloud [11]: A community cloud is established when multiple organizations collaborate to share cloud resources based on mutual interests, policies, and requirements. The cloud infrastructure can be managed by a third-party provider or jointly by community members. The primary advantages of this model include reduced costs due to resource sharing and enhanced security tailored to the needs of the community.

Hybrid Cloud [12]: A hybrid cloud is formed by integrating two or more distinct cloud environments—public, private, or community—while maintaining their individual characteristics. This model ensures interoperability and data portability through standardized or mutually agreed functionalities, allowing seamless communication between different cloud components.

5. ML-Centric Resource Management: Challenges and Approaches

This section explores challenges in ML-based resource management, discussing current solutions, their benefits, and limitations. Challenges are categorized into workload prediction, VM consolidation, thermal management, and resource provisioning.

5.1. Workload Prediction

5.1.1. ML for Energy Consumption Forecasting

Cloud providers typically estimate energy usage offline, but real-time predictions remain challenging due to dynamic workloads. A study by Reiss et al. [13] revealed that Google clusters utilize only 60% of CPU and 50% of memory on average. This suggests ensemble learning could enhance accuracy in cloud energy forecasting.

Subirats and Guitart [14] introduced an ensemble learning method combining moving average, exponential smoothing, linear regression, and double exponential smoothing to predict VM resource usage and improve energy efficiency. Their approach calculates the Mean Absolute Error (MAE) across iterations to select the best model. However, they overlooked key factors like Last-Level Cache (LLC) and disk throughput [15], which significantly impact energy consumption. Additionally, prediction accuracy varies based on workload types (interactive vs. batch), limiting generalization.

5.1.2. Performance and Online Workload Profiling

Cloud resource management research often neglects VM lifetime resource consumption. While offline workload profiling is impractical due to unavailable input workloads before production, online profiling is complex as it is difficult to determine when a VM exhibits representative behavior [16].

On Microsoft Azure, Bianchini et al. [16] developed an ML-based system that learns from historical data and predicts workload behavior in real-time, benefiting various resource managers. Their study revealed consistent CPU utilization peaks across VMs. However, they did not account for memory usage, which significantly impacts resource exhaustion. Additionally, their workload classification method using Extreme Gradient Boosting Tree (EGBT) did not address the challenge of distributed data centers with partially labeled datasets, which limits training effectiveness.

5.1.3. . Prediction Accuracy in Auto-Scaling Web Applications

Auto-scaling allocates resources dynamically, using either reactive or proactive strategies [17]. The reactive approach scales resources based on system events like CPU usage exceeding thresholds, while proactive scaling anticipates future demands. Traditional statistical models used in proactive scaling often struggle with prediction accuracy due to:

- Rule-based dependencies between variables, limiting flexibility.
- Poor scalability to high-dimensional data, as they rely on a limited set of attributes.

ML-based approaches can improve auto-scaling efficiency, but achieving high accuracy remains a challenge.

5.1.4. Training data

In modern cloud environments, virtual resources such as virtual CPUs (vCPUs) and memory (vRAMs) have a nonlinear resource demand, resulting in complex resource utilization behavior. As a result, optimization of virtual resource performance is required with this high amount of daily workload. Large corporations such as Amazon, Alibaba,

and others have occasionally failed due to a lack of resource management planning. As a result, predicting virtual resources (such as vCPU and vRAM) is a challenging task. Furthermore, resource forecasting presents some challenges: (1) The prediction of these resources should be dynamic to respond to changing workload patterns over time; (2) The data for training should be chosen in such a way that it has the most significant impact on the target variable so that the model can learn to predict it effectively.

5.2. Runtime VM Management

5.2.1. VM Consolidation and Resource Usage

VM consolidation aims to optimize energy efficiency by running more VMs on fewer hosts and shutting down idle ones. However, most methods rely solely on CPU utilization to detect overloaded hosts, leading to unnecessary VM migrations and power mode transitions. Additionally, neglecting future resource demands can result in overutilization of the destination host.

Haghshenas and Mohammadi [18] proposed an intelligent VM consolidation approach using historical data to predict resource utilization and optimize VM migration. They implemented a dynamic consolidation strategy using Linear Regression (LR) on real workload traces from PlanetLab VMs [19] within the CloudSim toolkit [20]. Their approach reduced energy consumption while accounting for time overheads. However, the reliance on LR, which requires extensive feature processing, could slow down response times in real-world deployments.

5.2.2. Multi-Dimensional Resource Management

Cloud data centers require dynamic resource provisioning to manage workloads effectively, but predicting demand for multiple resources (CPU, memory, disk, network) is complex. Since VM requests vary widely, forecasting resource needs accurately remains a challenge.

Ismaeel and Miri [21] this by categorizing VMs into clusters and applying Extreme Learning Machines (ELMs) for prediction. Their approach offers advantages such as:

- Faster training by optimizing predictor weights in a single step.
- Avoiding issues related to learning rates, stopping conditions, and local minima.

- Handling nonlinear processes better than LR.
- Using a single network per cluster to predict VM requests.
- Allowing customized prediction models for each cluster.

However, their method fixes the number of clusters at three using K-means clustering, which may limit adaptability in dynamic environments.

5.3. VM Placement

5.3.1. Cloud Network Traffic

VM placement strategies often allocate resources based on current CPU utilization, leading to inefficient resource use as workload demands fluctuate. Future resource estimation, including CPU and network bandwidth, is crucial for effective VM placement [22, 23]. High network traffic in cloud computing environments impacts VM migration time and can violate SLAs [24].

Shaw et al. [25] proposed a network-aware predictive VM placement heuristic that considers both CPU and network bandwidth demands. This approach enhances scheduling decisions and improves reliability by reducing energy consumption and SLA violations. However, it does not account for disk throughput, which also affects migration efficiency [26].

5.4. Thermal Management

5.4.1. Host Temperature

Managing host temperature in cloud data centers is critical, as heat dissipation increases cooling costs and leads to system failures. High temperatures create hotspots, impacting performance and reliability. Ilager et al. [27] introduced a thermal-aware predictive scheduling approach to minimize peak temperatures and energy consumption. Using sensor data from the University of Melbourne's private cloud, they trained ML models to predict temperature fluctuations and optimize VM migration for thermal management.

5.5. Resource Provisioning

5.5.1. SLA-Based VM Management

Over-provisioning is commonly used in data centers to prevent SLA violations during peak demand but leads to resource wastage and increased cooling costs [28, 29]. Dynamic resource provisioning has been explored to address this, but many approaches focus on specific applications rather than diverse workloads.

Garg et al. [30] proposed an SLA-aware resource management strategy that distinguishes between compute-intensive and transactional workloads. Using historical CPU utilization data and SLA penalties, they developed an artificial neural network (ANN) to predict CPU demand over two-hour intervals. However, their method struggles with high workload variability, sometimes deviating from actual values and failing to account for nonlinear workload behavior, which can impact QoS and energy efficiency [31].

5.5.2. QoS-Aware Resource Provisioning

Fluctuating application demands in cloud environments make static resource allocation inefficient, leading to either wasted resources or performance degradation. Dynamic resource provisioning is essential, with proactive strategies predicting future loads to ensure QoS compliance.

Calheiros et al. [32] developed an ARIMA-based workload prediction model to dynamically provision VMs based on expected demand while maintaining QoS parameters like response time and rejection rate. Their approach leveraged historical web request data [33], but the static time interval for VM deployment could create inefficiencies if the estimated time did not align with actual provisioning requirements, potentially affecting response times.

6. Future Research Directions

6.1. Workload Prediction

6.1.1. ML in Energy Consumption Prediction

Future research should consider non-linear relationships (e.g., polynomial, exponential) between system power metrics (CPU, memory, disk, network) and energy consumption. Instead of selecting the best model in ensemble learning, an optimized approach could aggregate weighted predictions based on mean absolute error. Adaptive real-time modeling of workload parameters could improve resource utilization. ML models must handle sudden resource usage changes to enhance prediction accuracy.

6.1.2. Performance and Online Profiling of Workload

Accurate workload estimation is crucial for intelligent resource management in complex cloud environments (Google, Microsoft, Amazon). Future research should explore advanced ML and DL models to improve workload predictions while minimizing computational complexity. Online profiling is essential to prevent resource exhaustion, and semi-supervised learning [34] could enhance classification accuracy in large-scale distributed data centers [35, 36].

6.1.3. Prediction Accuracy in Auto-Scaling Web Applications

ML models offer advantages in auto-scaling, such as learning from large datasets and adapting without explicit programming. However, processing redundant features increases computational overhead. Future research should focus on feature selection techniques like wrappers, filters, and embedded methods [37] to optimize speed-accuracy trade-offs in large datasets.

6.1.4. Time-Series Prediction

Developing a generalized ensemble framework for time-series workload forecasting remains a key research area. Novel architectures, including CNNs and attention mechanisms, can improve accuracy [38, 39]. Temporal Convolution Networks (TCNs) have shown promise for sequence modeling [40, 41] and could outperform RNNs in forecasting workload patterns.

6.1.5. Data Training

Hyperparameter optimization impacts ML training performance. Future research should explore optimization techniques such as Grid Search, Random Search, Bayesian Optimization, and Gradient-Based Optimization [42] to enhance model efficiency.

6.2. Runtime VM Management

6.2.1. Multi-Resource Usage in VM Consolidation

Future work should integrate CPU, memory, and bandwidth metrics to identify overloaded hosts accurately [43, 44]. DL models like LSTM and GRU could optimize VM consolidation by reducing training overhead and improving efficiency in large-scale data centers.

6.2.2. Multi-Dimensional Resource Requirement

Instead of fixed clustering methods (e.g., K-means), ensemble clustering techniques [45, 46] can improve VM classification. Advanced clustering methods can enhance accuracy, reduce time complexity, and optimize resource consumption.

6.2.3. Energy Metering at Software Level

Clustering analysis could categorize VMs based on energy consumption levels instead of relying on difficult-to-measure VM-level power metrics [47, 48]. Feature selection methods like ChiSquare Score and Fisher Score [49] could improve clustering efficiency.

6.2.4. Usage Level Management

Future research should develop dynamic resource utilization thresholds for VM migration, preventing unnecessary migrations when short-term fluctuations occur. Adaptive strategies can enhance consolidation efficiency.

6.3. VM Placement

6.3.1. Cloud Network Traffic

VM placement strategies should integrate disk throughput as a key factor alongside CPU and network bandwidth [50-53]. Optimizing disk I/O could prevent tail latency issues, which affect online service performance and SLA compliance [54-56].

6.4. Thermal Management

6.4.1. Host Temperature

Predicting CPU temperature in advance can enhance thermal management strategies and reduce cooling costs. Future research should focus on CPU load estimation to prevent overheating rather than reactively cooling overloaded hosts [57]. Using GRU-based single-feature models could reduce algorithm complexity and improve scheduling efficiency.

These research directions aim to enhance cloud resource management through advanced ML and DL techniques while improving efficiency, scalability, and QoS.

7. Conclusion

Cloud computing systems are vast, highly interconnected, and resource-intensive, making effective resource management a complex challenge. Traditional rule-based and heuristic resource management strategies struggle with scalability, heterogeneity, and dynamic workload demands. Data-driven AI techniques, particularly machine learning (ML), have emerged as powerful tools for optimizing resource allocation, workload prediction, and energy consumption management.

This paper explores the key challenges of resource management in cloud environments and examines various ML-based solutions, highlighting their advantages and limitations. Recent studies have increasingly leveraged ML models to enhance workload forecasting, optimize energy usage, and improve overall efficiency. Different ML techniques are applied to specific problems, each with varying effectiveness depending on the use case. This paper also identifies future research directions, aiming to refine ML techniques for cloud resource management. Advanced ML approaches, such as deep learning and reinforcement learning, hold significant potential for intelligent resource optimization. By adopting these cutting-edge methods, cloud computing systems can achieve higher efficiency, better scalability, and improved performance in handling complex tasks. This study serves as a valuable reference for researchers seeking to explore and enhance ML applications in cloud resource management.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest is to be disclosed.

References

[1] Global digital population as of July 2020. https://www.statista.com/statistics/617136/digitalpopulation-worldwide/

[2] Lakshmi Devasena C (2014) Impact study of cloud computing on business development. Oper Res Appl Int J (ORAJ) 1:1–7

[3] Tsakalidou, V.N., Mitsou, P., Papakostas, G.A. (2022). Machine Learning for Cloud Resources Management—An Overview. In: Smys, S., Lafata, P., Palanisamy, R., Kamel, K.A. (eds) Computer Networks and Inventive Communication Technologies. Lecture Notes on Data Engineering and Communications Technologies, vol 141. Springer, Singapore. https://doi.org/10.1007/978-981-19-3035-5_67

[4] Mell, Peter, 2011. The NIST definition of cloud computing. In N. I. O. S. A. Technology (Ed.): U.S. Department of Commerce.

[5] Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine learning (ML)-centric resource management in cloud computing: A review and future directions. Journal of Network and Computer Applications, 204, 103405.

[6] Piraghaj, Sareh Fotuhi, Dastjerdi, Amir Vahid, Calheiros, Rodrigo N, Buyya, Rajkumar, 2017. A survey and taxonomy of energy efficient resource management techniques in platform as a service cloud. In: Handbook of Research on End-to-End Cloud Computing Architecture Design. IGI Global, pp. 410–454.

[7] Jula, Amin, Sundararajan, Elankovan, Othman, Zalinda, 2014. Cloud computing service composition: A systematic literature review. Expert Syst. Appl. 41 (8), 3809–3824.

[8] Whaiduzzaman, Md, Sookhak, Mehdi, Gani, Abdullah, Buyya, Rajkumar, 2014. A survey on vehicular cloud computing. J. Netw. Comput. Appl. 40, 325–344.

[9] Toosi, Adel Nadjaran, Calheiros, Rodrigo N., Buyya, Rajkumar, 2014. Interconnected cloud computing environments: Challenges, taxonomy, and survey. ACM Comput. Surv. 47 (1), 1–47.

[10] Jadeja, Yashpalsinh, Modi, Kirit, 2012. Cloud computing-concepts, architecture and challenges. In: 2012 International Conference on Computing, Electronics and Electrical Technologies. ICCEET, IEEE, pp. 877–880.

[11] Dillon, Tharam, Wu, Chen, Chang, Elizabeth, 2010. Cloud computing: issues and challenges. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications. Ieee, pp. 27–33.

[12] Tuli, Shreshth, Sandhu, Rajinder, Buyya, Rajkumar, 2020. Shared data-aware dynamic resource provisioning and task scheduling for data intensive applications on hybrid clouds using aneka. Future Gener. Comput. Syst. 106, 595–606.

[13] Reiss, Charles, Wilkes, John, Hellerstein, Joseph L., 2011. Google Cluster-Usage Traces: Format+ Schema. White Paper, Google Inc. pp. 1–14.

[14] Subirats, Josep, Guitart, Jordi, 2015. Assessing and forecasting energy efficiency on cloud computing platforms. Future Gener. Comput. Syst. 45, 70–94.

[15] Sayadnavard, M. H., Haghighat, A. T., & Rahmani, A. M. (2022). A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers. Engineering science and technology, an International Journal, 26, 100995.

[16] Bianchini, Ricardo, Fontoura, Marcus, Cortez, Eli, Bonde, Anand, Muzio, Alexandre, Constantin, Ana-Maria, Moscibroda, Thomas, Magalhaes, Gabriel, Bablani, Girish, Russinovich, Mark, 2020. Toward ML-centric cloud platforms. Commun. ACM 63(2), 50–59

[17] Persico, Valerio, Grimaldi, Domenico, Pescape, Antonio, Salvi, Alessandro, Santini, Stefania, 2017. A fuzzy approach based on heterogeneous metrics for scaling out public clouds. IEEE Trans. Parallel Distrib. Syst. 28 (8), 2117–2130

[18] Haghshenas, Kawsar, Mohammadi, Siamak, 2020. Prediction-based underutilized and destination host selection approaches for energy-efficient dynamic VM consolidation in data centers. J. Supercomput. 1–18.

[19] Chun, Brent, Culler, David, Roscoe, Timothy, Bavier, Andy, Peterson, Larry, Wawrzoniak, Mike, Bowman, Mic, 2003. Planetlab: an overlay testbed for broad-coverage services. ACM SIGCOMM Comput. Commun. Rev. 33 (3), 3–12

[20] Calheiros, Rodrigo N, Ranjan, Rajiv, Beloglazov, Anton, De Rose, César AF, Buyya, Rajkumar, 2011. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. - Pract.Exp. 41 (1), 23–50.

[21] Ismaeel, Salam, Miri, Ali, 2015. Using ELM techniques to predict data centre VM requests. In: 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing. IEEE, pp. 80–86.

[22] Genez, Thiago AL, Bittencourt, Luiz F, da Fonseca, Nelson LS, Madeira, Edmundo RM, 2015. Estimation of the available bandwidth in inter-cloud links for task scheduling in hybrid clouds. IEEE Trans. Cloud Comput. 7 (1), 62–74.

[23] Duggan, Martin, Duggan, Jim, Howley, Enda, Barrett, Enda, 2017. A network aware approach for the scheduling of virtual machine migration during peak loads. Cluster Comput. 20 (3), 2083–2094

[24] Verma, Akshat, Ahuja, Puneet, Neogi, Anindya, 2008. pMapper: power and migration cost aware application placement in virtualized systems. In: ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing. Springer, pp. 243–264.

[25] Shaw, Rachael, Howley, Enda, Barrett, Enda, 2019. An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. Simul. Model. Pract. Theory 93, 322–342

[26] Brewer, Eric, Ying, Lawrence, Greenfield, Lawrence, Cypher, Robert, T'so, Theodore, 2016. Disks for data centers.

[27] Ilager, S., Ramamohanarao, K., Buyya, R., 2021. Thermal prediction for efficient energy management of clouds using machine learning. IEEE Trans. Parallel Distrib. Syst. 32 (5), 1044–1056. http://dx.doi.org/10.1109/TPDS.2020.3040800.

[28] Reiss, Charles, Tumanov, Alexey, Ganger, Gregory R, Katz, Randy H, Kozuch, Michael A, 2012. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In: Proceedings of the Third ACM Symposium on Cloud Computing. pp. 1–13.

[29] Sun, Xiang, Ansari, Nirwan, Wang, Ruopeng, 2016. Optimizing resource utilization of a data center. IEEE Commun. Surv. Tutor. 18 (4), 2822–2846.

[30] Garg, Saurabh Kumar, Toosi, Adel Nadjaran, Gopalaiyengar, Srinivasa K, Buyya, Rajkumar, 2014. SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter. J. Netw. Comput. Appl. 45, 108–120.

[31] Kumar, Jitendra, Singh, Ashutosh Kumar, Buyya, Rajkumar, 2020. Self directed learning based workload forecasting model for cloud resource management. Inform. Sci. 543, 345–366

[32] Calheiros, Rodrigo N, Masoumi, Enayat, Ranjan, Rajiv, Buyya, Rajkumar, 2014. Workload prediction using ARIMA model and its impact on cloud applications' QoS. IEEE Trans. Cloud Comput. 3 (4), 449–458.

[33] Amekraz, Zohra, Hadi, Moulay Youssef, 2018. Higher order statistics based method for workload prediction in the cloud using ARMA model. In: 2018 International Conference on Intelligent Systems and Computer Vision. ISCV, IEEE, pp. 1–5.

[34] Zhu, Xiaojin, Goldberg, Andrew B., 2009. Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. 3 (1), 1–130.

[35] Cortez, Eli, Bonde, Anand, Muzio, Alexandre, Russinovich, Mark, Fontoura, Marcus, Bianchini, Ricardo, 2017. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In: Proceedings of the 26th Symposium on Operating Systems Principles. pp. 153–167.

[36] Bianchini, Ricardo, Fontoura, Marcus, Cortez, Eli, Bonde, Anand, Muzio, Alexandre, Constantin, Ana-Maria, Moscibroda, Thomas, Magalhaes, Gabriel, Bablani, Girish, Russinovich, Mark, 2020. Toward ML-centric cloud platforms. Commun. ACM 63(2), 50–59.

[37] Majeed, Abdul, 2019. Improving time complexity and accuracy of the machine learning algorithms through selection of highly weighted top k features from complex datasets. Ann. Data Sci. 6 (4), 599–621.

[38] Lai, Guokun, Chang, Wei-Cheng, Yang, Yiming, Liu, Hanxiao, 2018. Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 95–104.

[39] Shih, Shun-Yao, Sun, Fan-Keng, Lee, Hung-yi, 2019. Temporal pattern attention for multivariate time series forecasting. Mach. Learn. 108 (8), 1421–1441.

[40] Borovykh, Anastasia, Bohte, Sander, Oosterlee, Cornelis W., 2017. Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691

[41] Bai, Shaojie, Kolter, J. Zico, Koltun, Vladlen, 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[42] Feurer, Matthias, Hutter, Frank, 2019. Hyperparameter optimization. In: Automated Machine Learning. Springer, Cham, pp. 3–33.

[43] Nguyen, Trung Hieu, Di Francesco, Mario, Yla-Jaaski, Antti, 2017. Virtual machine consolidation with multiple usage prediction for energy-efficient cloud data centers. IEEE Trans. Serv. Comput.

[44] Abdelsamea, Amany, El-Moursy, Ali A, Hemayed, Elsayed E, Eldeeb, Hesham, 2017. Virtual machine consolidation enhancement using hybrid regression algorithms. Egypt. Inform. J. 18 (3), 161–170.

[45] Alqurashi, Tahani, Wang, Wenjia, 2019. Clustering ensemble method. Int. J. Mach. Learn. Cybern. 10 (6), 1227–1246

[46] Boongoen, Tossapon, Iam-On, Natthakan, 2018. Cluster ensembles: A survey of approaches with recent extensions and applications. Comp. Sci. Rev. 28, 1–25

[47] Kansal, Aman, Zhao, Feng, Liu, Jie, Kothari, Nupur, Bhattacharya, Arka A, 2010. Virtual machine power metering and provisioning. In: Proceedings of the 1st ACM Symposium on Cloud Computing. pp. 39–50.

[48] Zhao-Hui, Y., Qin-Ming, J., 2012. Power management of virtualized cloud computing platform. Chinese J. Comput. 6, 015.

[49] Vora, Suchi, Yang, Hui, 2017. A comprehensive study of eleven feature selection algorithms and their impact on text classification. In: 2017 Computing Conference. IEEE, pp. 440–449.

[50] Shaw, Rachael, Howley, Enda, Barrett, Enda, 2019. An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. Simul. Model. Pract. Theory 93, 322–342.

[51] Md Bahar Uddin, Md. Hossain and Suman Das, "Advancing manufacturing sustainability with industry 4.0 technologies", International Journal of Science and Research Archive, 2022, 06(01), 358-366.

[52] Bharati, S., Rahman, M. A., & Podder, P. (2018, September). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (pp. 581-584). IEEE.

[53] Bharati, S., Mondal, M. R. H., Podder, P., & Prasath, V. S. (2022). Federated learning: Applications, challenges and future directions. International Journal of Hybrid Intelligent Systems, 18(1-2), 19-35.

[54] Bharati, S., Mondal, M. R. H., Podder, P., & Prasath, V. S. (2022). Federated learning: Applications, challenges and future directions. International Journal of Hybrid Intelligent Systems, 18(1-2), 19-35.

[55] Bharati, S., Podder, P., Mondal, M. R. H., Podder, P., & Kose, U. (2022). A review on epidemiology, genomic characteristics, spread, and treatments of COVID-19. Data Science for COVID-19, 487-505.

[56] Brewer, Eric, Ying, Lawrence, Greenfield, Lawrence, Cypher, Robert, T'so, Theodore, 2016. Disks for data centers.

[57] Ilager, S., Ramamohanarao, K., Buyya, R., 2021. Thermal prediction for efficient energy management of clouds using machine learning. IEEE Trans. Parallel Distrib. Syst. 32 (5), 1044–1056. http://dx.doi.org/10.1109/TPDS.2020.3040800.