© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722726

AI-Driven Capacity Forecasting and Resource Optimization in Enterprise Cloud Infrastructure in Saudi Arabia

MUNEER SHAIK

Abstract- ***Enterprise cloud environments generate continuous streams of utilization data, yet many organizations still plan capacity through static thresholds, periodic reports, and reactive expansion. This review examines how artificial intelligence (AI) can convert infrastructure telemetry into forward-looking capacity forecasts and resource decisions for enterprise cloud platforms, with specific attention to Saudi Arabia. A structured review of recent peer-reviewed literature (2020-2025) and official Saudi digital-transformation sources was undertaken, covering workload prediction, deep learning, reinforcement learning, autoscaling, resource allocation, energy efficiency, and AIOps. The literature indicates that recurrent, attention-based, hybrid, and uncertainty-aware forecasting models can improve anticipation of CPU, memory, storage, and workload demand, while reinforcement-learning approaches can translate predictions into scaling, placement, and scheduling actions. However, accuracy alone is insufficient: useful enterprise systems require data quality controls, application context, explainability, service-level safeguards, and human-governed automation. For Saudi organizations, these capabilities align with Cloud First and Vision 2030 objectives by supporting resilient digital services, efficient infrastructure investment, and scalable modernization. The review proposes a layered implementation framework combining observability, predictive models, policy engines, and controlled automation, and identifies research priorities for multi-resource forecasting, uncertainty quantification, cross-domain transfer, and sovereign-cloud operations.***

Keywords: artificial intelligence; capacity forecasting; cloud computing; resource optimization; AIOps; autoscaling; Saudi Arabia; Vision 2030

I. INTRODUCTION

Cloud infrastructure has moved from a supporting utility to a core platform for enterprise applications, analytics, virtual desktops, digital public services, and business-critical systems. This transition increases the operational cost of inaccurate capacity decisions. Over-provisioning ties up capital and energy in idle compute, memory, and storage; under-provisioning creates contention, response-time degradation, failed jobs, or service-level agreement (SLA) violations. The problem is particularly difficult in virtualized and hyperconverged environments because demand changes at several layers simultaneously: applications generate time-varying requests, virtual machines share physical hosts, storage latency influences application behavior, and maintenance or disaster-recovery events can suddenly shift load.

Recent cloud-resource research therefore increasingly treats capacity planning as a prediction-and-control problem rather than a periodic reporting exercise. Reviews of machine-learning-centred resource management identify workload estimation, task scheduling, virtual-machine consolidation, resource optimization, and energy optimization as connected functions of data-driven cloud management. Comparative work also shows that no single forecasting family is universally superior across heterogeneous workload traces, making model selection, evaluation windows, and data characteristics important. Deep-learning studies have consequently explored recurrent networks, convolutional models, attention mechanisms, generative methods, and hybrid architectures for nonlinear workload behaviour.

For Saudi Arabia, the subject has strategic relevance. National digital-transformation programs promote cloud adoption, e-government integration, and advanced digital infrastructure, while the Kingdom's Cloud First policy encourages government entities to consider cloud options for new IT investments. As more critical workloads move to shared, hybrid, and cloud-oriented platforms, capacity decisions need to

become faster, more evidence-based, and more resilient. This paper reviews the role of AI-driven capacity forecasting in that transition and develops an enterprise-oriented framework for applying prediction to resource optimization without treating automation as a substitute for governance.

II. BACKGROUND AND LITERATURE REVIEW

Traditional capacity management is usually built around historical utilization, fixed headroom percentages, threshold alarms, and administrator judgement. These techniques remain useful for baseline governance, but they respond poorly to bursty, nonstationary, or multi-resource demand. Cloud workload studies demonstrate that CPU, memory, storage, and request rates can contain periodicity, abrupt spikes, long-term dependencies, and cross-metric relationships. Bi et al. combined preprocessing with bidirectional and grid long short-term memory networks for workload and resource prediction, while Malik et al. evaluated evolutionary and machine-learning methods for CPU and memory utilization. Dogani et al. used wavelet decomposition with BiGRU and attention to address nonstationary host load, and a related multivariate study combined CNN, GRU, and attention to predict multiple resource dimensions.

Attention-based and generative approaches extend this direction. Patel and Bedi's MAG-D architecture emphasized multivariate attention for cloud workload forecasting, while Maiyza et al. investigated a hybrid generative-adversarial approach. Studies from King Abdulaziz University examined both deep-learning workload prediction and deep reinforcement learning in federated cloud settings. Karthikeyan and Sundaravadivazhagan similarly linked AI-augmented prediction with sustainable cloud data-centre operation. Across these studies, a consistent theme emerges: forecasting is most valuable when it is connected to an operational decision, such as scaling capacity before demand arrives or moving workload before a host becomes saturated.

Resource optimization research addresses that second step. Reinforcement learning (RL) is attractive because a controller can learn a policy that balances competing goals such as performance, cost, utilization, fault tolerance, and energy. Surveys of RL-based autoscaling describe adaptive scaling as a sequential decision problem under uncertainty. Xue et al. proposed meta-reinforcement learning for predictive autoscaling, and Abdullah et al. addressed workload bursts in containerized microservices. Other work applies RL or multi-agent learning to time-varying scheduling, resource allocation, and task placement. Energy-efficiency reviews further show that virtualization, workload placement, migration, and application-level controls can materially influence data-centre power demand.

AIOps broadens the objective beyond forecasting a single metric. It combines operational telemetry with AI to support anomaly detection, failure prediction, root-cause analysis, and automated action. A systematic review by Diaz-de-Arcaya et al. highlights both the opportunities and organizational challenges of AIOps and MLOps adoption. More recent uncertainty-aware workload forecasting also shows why a point estimate is not enough: resource managers benefit from knowing how confident the model is, especially where under-provisioning carries high service risk. Forecast horizon is equally important. A five-minute prediction can support autoscaling, whereas a weekly or monthly forecast is more relevant to storage expansion, hardware procurement, or reserved-cloud commitments. Enterprise capacity management therefore benefits from a hierarchy of models rather than a single predictor: short-horizon models for operational control, medium-horizon models for maintenance and workload movement, and longer-horizon models for financial and architectural planning. Evaluation should consequently include not only MAE or RMSE but also missed-peak rate, over-provisioning cost, SLA impact, and forecast stability across changing workload regimes.

III. RESEARCH AIM, OBJECTIVES AND QUESTIONS

The aim of this review is to assess how AI-driven capacity forecasting can improve resource utilization, resilience, and operational efficiency in enterprise

cloud infrastructure, and to interpret those findings for Saudi Arabia's digital-transformation context.

The objectives are to: (1) examine current AI techniques for cloud workload and capacity forecasting; (2) evaluate how forecasts can support resource allocation, autoscaling, health monitoring, and service continuity; (3) identify implementation risks and governance requirements; and (4) propose a practical framework for Saudi enterprise environments.

The review addresses three questions: RQ1: How can AI improve capacity forecasting in enterprise cloud infrastructure? RQ2: How can prediction be translated into safer and more efficient resource optimization? RQ3: What technical and governance factors are most important for adoption in Saudi enterprises?

IV. REVIEW METHODOLOGY

A structured narrative review was used because the topic spans predictive modelling, cloud operations, optimization, and national digital-infrastructure policy. The search strategy focused on Scopus- and Web of Science-indexed journals and major computing libraries, including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Wiley. Search combinations included "cloud workload prediction," "resource utilization forecasting," "AI capacity planning," "predictive autoscaling," "reinforcement learning cloud resource allocation," "AIOps cloud operations," and "cloud data-center energy efficiency."

Peer-reviewed studies published mainly between 2020 and 2025 were prioritized. Papers were included when they addressed forecasting of cloud workload or resource demand, prediction-driven resource control, autoscaling, task or VM allocation, or operational AI. Studies focused only on unrelated edge-device optimization or without a clear cloud-resource-management contribution were excluded. Twenty-three peer-reviewed sources were synthesized together with two official Saudi sources covering Cloud First and Vision 2030 digital transformation. Findings were organized thematically around forecasting, optimization, operational intelligence, resilience, and Saudi implementation. Priority was given to sources with a verifiable DOI or an official Saudi-government publication origin, and duplicate or weakly documented items were not used as evidence. The review is interpretive rather than meta-analytic because the selected studies use different datasets, forecast horizons, error metrics, and optimization objectives. This heterogeneity prevents a defensible pooled accuracy estimate; instead, the synthesis focuses on recurring design principles, operational trade-offs, and areas where multiple studies converge.

V. THEMATIC FINDINGS AND ANALYSIS

5.1 Predictive Capacity Modelling

AI improves capacity forecasting by learning temporal and cross-resource patterns that are difficult to encode as static thresholds. Recurrent architectures such as LSTM and GRU are frequently used because infrastructure telemetry is sequential. Comparative studies indicate that deep models can capture nonlinear workload behaviour, although performance varies by dataset and forecast horizon. Hybrid models attempt to improve stability by combining signal decomposition, convolution, attention, or feature-selection stages before recurrent prediction. Generative and evolutionary methods provide additional ways to model irregularity or optimize predictor structure. For enterprise use, the most important design choice is not the model label but the prediction target. Forecasting CPU alone can miss storage latency, memory pressure, or application-level constraints; multi-resource forecasting is therefore more useful for real capacity planning. A practical feature set can combine utilization percentiles, allocation-versus-consumption ratios, queue depth, storage IOPS and latency, network throughput, VM density, cluster free capacity, replication state, and application transaction volume. Calendar and operational events—month-end processing, software releases, backup windows, patching, or planned failovers—can also improve prediction when represented explicitly. This matters because many apparent "anomalies" are predictable business events that conventional threshold systems treat as surprises. Models should therefore be trained

against both technical telemetry and workload context, then validated separately on normal periods and peak periods to avoid reporting good average accuracy while failing during the exact conditions that capacity planning is intended to protect.

5.2 Resource Optimization and Autoscaling

Forecasts create value when they trigger controlled decisions before congestion occurs. Predictive autoscaling can add or remove resources using expected demand rather than waiting for an alarm. RL methods are particularly relevant when the controller must learn trade-offs among SLA risk, utilization, cost, and scaling frequency. Burst-aware methods also show that average demand is an incomplete signal for public-facing services. Scheduling and multi-agent studies extend the same logic to placement of jobs and VMs across heterogeneous hosts. In practice, an enterprise policy engine should combine model output with hard guardrails: minimum redundancy, reserved headroom for critical services, maintenance windows, licensing limits, data-locality rules, and change-management controls. AI should recommend or execute within those boundaries rather than optimize a single utilization metric in isolation. In virtualized and hyperconverged infrastructure, recommended actions can include VM right-sizing, adding hosts to a cluster, rebalancing workloads, expanding storage pools, adjusting reservations, or scheduling live migration. The cost of each action differs: vertical scaling may require restart, host addition has procurement or subscription implications, and migration can consume network and storage bandwidth. A mature optimizer should therefore rank actions by expected benefit, operational disruption, reversibility, and risk. Where confidence is low, the safest output may be an operator recommendation rather than an automated change.

5.3 Infrastructure Health Scoring and AIOps

Capacity data becomes more actionable when combined with health and service context. A server at 75% CPU may be safe if its workload is stable, but risky if storage latency is rising, memory reclamation is active, replication is behind, or a failover event is expected. AIOps therefore supports a composite operational view that can combine metrics, logs, topology, incidents, and application dependencies. The resulting health score can prioritize which forecast requires action. This is particularly useful in large estates where administrators cannot manually inspect thousands of signals. However, health scoring must remain explainable: operators should be able to see which metrics influenced a risk score, how the forecast changed, and which policy caused a recommended action. A useful health score is therefore not a mysterious aggregate number. It should expose component scores for compute saturation, memory pressure, storage performance, capacity runway, replication or backup state, and service dependency. Trend direction is also important: a moderate utilization level that is rising quickly can deserve higher priority than a higher but stable level. By connecting forecast confidence and health severity, operations teams can triage capacity risks according to both probability and business impact.

5.4 Resilience and Service Continuity

Prediction also supports resilience. Capacity planning for disaster recovery is not simply a question of average utilization; the recovery site must absorb displaced workloads while maintaining priority services. Forecasts can estimate expected failover demand, identify resource shortfalls, and help stage migrations or expansions in advance. Multi-site environments further benefit from workload placement decisions that account for available capacity, network conditions, and service criticality. The literature on reinforcement learning and energy-aware scheduling demonstrates that placement policies can optimize multiple objectives, but enterprise resilience requires conservative constraints around fault domains and recovery point or recovery time objectives. The strongest use case is therefore human-governed predictive operations: AI detects emerging capacity risk early, proposes a response, and automation executes only within approved reliability limits. This design is especially relevant to multi-site enterprise platforms. A planned cross-cluster migration can be evaluated against destination headroom before movement begins; a disaster-recovery test can be modelled as a temporary demand surge; and maintenance can be sequenced so that remaining hosts retain safe capacity. Prediction can

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722726

also improve recovery exercises by identifying whether the secondary site can sustain not only booted workloads but their expected peak demand after users reconnect. These examples show that capacity forecasting is part of resilience engineering, not merely cost optimization.

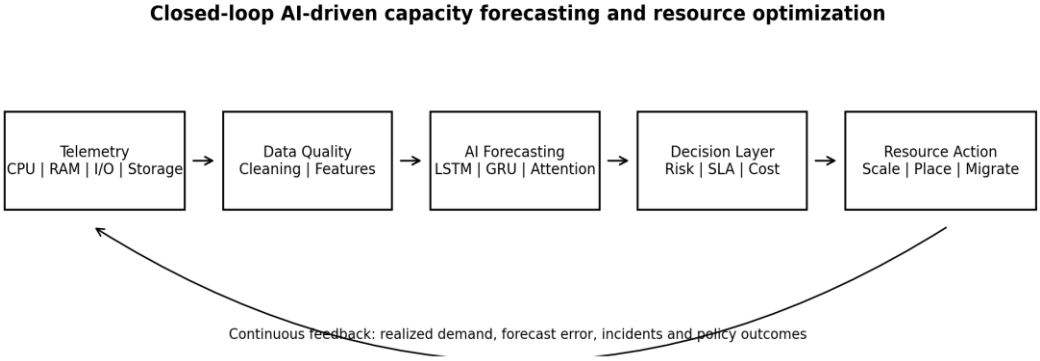

Figure 1. Closed-loop AI-driven capacity forecasting and resource optimization model.

Table 1. Traditional versus AI-driven enterprise capacity management

Dimension	Traditional approach	AI-driven approach
Primary trigger	Threshold breach, periodic review	Forecasted demand, uncertainty and risk
Data used	Recent utilization and static headroom	Multi-period telemetry, topology and application context
Planning horizon	Reactive or fixed planning cycle	Minutes to months, depending on use case
Resource action	Manual expansion or static reservation	Right-sizing, autoscaling, placement, migration
Main benefit	Simplicity and transparent rules	Earlier action and better matching of capacity to demand
Main risk	Late response or persistent over-provisioning	Model error, drift, automation without adequate guardrails

VI. DISCUSSION

The reviewed evidence supports a shift from reactive monitoring to predictive resource management, but it also shows why "AI-driven" should not be interpreted as fully autonomous infrastructure. Forecasting models are sensitive to training data, workload drift, sampling intervals, and sudden events. A model trained on routine business cycles may perform poorly during a major system rollout, cyber incident, disaster-recovery exercise, or seasonal surge. Cross-provider transfer can also be weak when workload distributions differ, which reinforces the need for local validation and uncertainty estimates.

A second issue is objective design. Resource utilization is not the same as business value. Driving a host toward very high utilization can reduce waste but may also increase queuing, reduce recovery headroom, and amplify failure impact. The appropriate target is a risk-adjusted operating range that reflects application criticality. This suggests a two-layer architecture: predictive models estimate future demand and uncertainty, while a policy layer decides what action is permissible. The policy layer can encode SLA class, redundancy, cost, energy, security, and maintenance constraints. Reinforcement learning may improve these decisions, but its reward function must reflect enterprise priorities rather than a narrow technical metric.

Third, observability quality is foundational. Missing telemetry, inconsistent naming, stale topology data, or unrecorded configuration changes can make sophisticated models unreliable. Organizations should therefore treat data engineering, asset context, and model monitoring as part of capacity management. The AIOps literature similarly emphasizes that operational AI requires integration across technical and organizational processes, not only model development. Another implication is that model performance itself must become an operational metric. Forecast error, calibration, false alarms, missed peaks, and the outcomes of automated actions should be logged and reviewed like any other service KPI. If error increases after a platform migration or application change, the system should reduce automation authority until the model is retrained and revalidated. This creates a controlled path from advisory AI to limited automation and, only where evidence supports it, to higher levels of autonomous operation. Figure 1 summarizes this closed-loop view.

VII. IMPLICATIONS FOR SAUDI ARABIA AND VISION 2030

Saudi Arabia's digital-transformation agenda creates a strong operational case for predictive cloud management. The Cloud First policy is intended to accelerate consideration of cloud solutions for government IT investment, while the National Transformation Program promotes digital infrastructure, e-government, and adoption of advanced technologies. As these initiatives expand the scale and criticality of digital services, infrastructure teams need methods that can grow capacity without simply maintaining large static reserves.

AI-driven forecasting can contribute in four practical ways. First, it can improve investment timing by showing when compute, memory, or storage capacity is likely to become constrained. Second, it can improve service continuity by identifying capacity risks before user-facing degradation occurs. Third, it can support hybrid and sovereign-cloud strategies by informing workload placement across approved environments. Fourth, it can reduce avoidable resource and energy waste by matching provisioned capacity more closely to observed and expected demand.

For Saudi enterprises, implementation should begin with high-value, well-instrumented platforms rather than estate-wide autonomy. Candidate environments include ERP, healthcare, virtual desktop, analytics, and other mission-critical systems with clear SLA and utilization history. Governance should align prediction-driven actions with cybersecurity requirements, data classification, change control, and disaster-recovery policy. Figure 2 presents a layered implementation model in which national and enterprise governance sits above observability, AI forecasting, orchestration, and measurable business outcomes. A staged Saudi adoption roadmap can begin with read-only forecasting dashboards, then progress to recommendation mode, approval-based automation, and finally low-risk autonomous actions. Each stage should have measurable exit criteria such as forecast accuracy across peak periods, reduction in emergency capacity incidents, acceptable false-alarm rates, successful rollback tests, and documented cybersecurity approval. This staged approach allows organizations to capture early value while maintaining the governance expected for critical national and enterprise systems.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722726

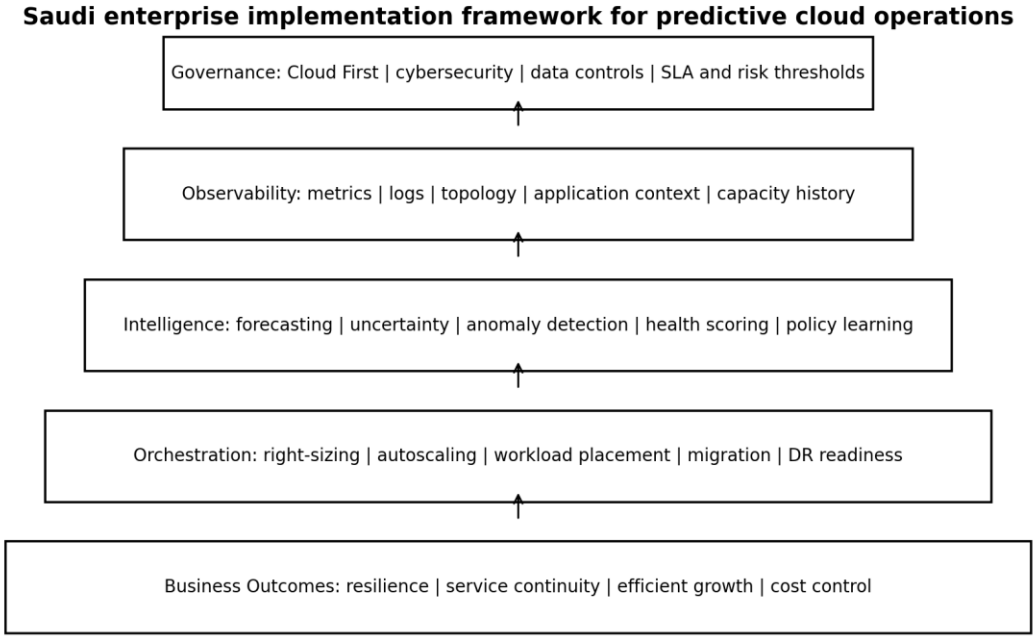

Figure 2. Layered Saudi enterprise framework for predictive cloud operations.

Table 2. AI techniques, operational uses and controls

Technique	Typical prediction/decision	Enterprise use	Required control
LSTM / GRU	Time-series workload and utilization	Capacity forecasting, trend anticipation	Back-testing, drift monitoring
CNN + attention / hybrid models	Multivariate and nonlinear patterns	CPU-memory-storage forecasting	Feature quality and horizon validation
Reinforcement learning	Sequential resource action	Autoscaling, placement, scheduling	Reward design, hard SLA guardrails
Bayesian / uncertainty-aware models	Forecast plus confidence	Risk-aware provisioning	Calibrated uncertainty thresholds
AIOps / health scoring	Risk, anomalies, incident context	Prioritization and proactive operations	Explainability and human approval for high-risk actions

VIII. CHALLENGES AND FUTURE RESEARCH

Several gaps remain. Multi-resource forecasting is still less mature than single-metric prediction, even though CPU, memory, storage, network, and application demand interact. Future work should evaluate joint forecasts and causal relationships rather than independent time series. Uncertainty-aware models also deserve greater attention because operators need confidence bounds, not only point forecasts. Another priority is concept drift: cloud workloads change after application releases, migrations, new business processes, or policy changes, so models require continuous evaluation and retraining.

For Saudi environments, research should also examine prediction across sovereign, private, and commercial cloud boundaries; Arabic and bilingual operational knowledge bases for AIOps; privacy-preserving model training; and the effect of data-residency constraints on workload placement.

Finally, explainable AI should be treated as an operational requirement. A capacity recommendation is more likely to be trusted when an engineer can trace it to recent demand, forecast confidence, service criticality, and the policy rule that produced the action.

IX. CONCLUSION

AI-driven capacity forecasting provides a credible path from periodic capacity reporting toward proactive cloud operations. The literature shows that recurrent, attention-based, hybrid, and uncertainty-aware models can anticipate complex workload patterns, while reinforcement learning and related optimization methods can convert forecasts into scaling, scheduling, and placement decisions. Their enterprise value, however, depends on more than prediction accuracy. Reliable deployment requires high-quality observability, application context, explicit SLA and risk constraints, explainable decisions, and human-governed automation.

For Saudi Arabia, these capabilities are consistent with the Kingdom's Cloud First and Vision 2030 digital-transformation direction. A practical adoption model is to begin with mission-critical, well-instrumented services; forecast multiple resource dimensions; attach uncertainty to predictions; and use a policy engine to control automation. This approach can improve resilience, reduce avoidable over-provisioning, and support scalable digital infrastructure while preserving governance. Future research should focus on multi-resource and cross-cloud prediction, model drift, uncertainty, explainability, and Saudi-specific sovereign-cloud operating conditions.

REFERENCES

[1] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," Journal of Network and Computer Applications, vol. 204, 103405, 2022. https://doi.org/10.1016/j.jnca.2022.103405

[2] D. Saxena, J. Kumar, A. K. Singh, and S. Schmid, "Performance analysis of machine learning centered workload prediction models for cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313–1330, 2023. https://doi.org/10.1109/TPDS.2023.3240567

[3] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, and A. S. Al-Malaise Al-Ghamdi, "Technical study of deep learning in cloud computing for accurate workload prediction," Electronics, vol. 12, no. 3, 650, 2023. https://doi.org/10.3390/electronics12030650

[4] J. Bi, S. Li, H. Yuan, and M. C. Zhou, "Integrated deep learning method for workload and resource prediction in cloud systems," Neurocomputing, vol. 424, pp. 35–48, 2021. https://doi.org/10.1016/j.neucom.2020.11.011

[5] S. Malik, M. Tahir, M. Sardaraz, and A. Alourani, "A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques," Applied Sciences, vol. 12, no. 4, 2160, 2022. https://doi.org/10.3390/app12042160

[6] J. Dogani, F. Khunjush, and M. Seydali, "Host load prediction in cloud computing with Discrete Wavelet Transformation (DWT) and Bidirectional Gated Recurrent Unit (BiGRU) network," Computer Communications, vol. 198, pp. 157–174, 2023. https://doi.org/10.1016/j.comcom.2022.11.018

[7] J. Dogani, F. Khunjush, M. R. Mahmoudi, and M. Seydali, "Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism," The Journal of Supercomputing, vol. 79, 2023. https://doi.org/10.1007/s11227-022-04782-z

[8] Y. S. Patel and J. Bedi, "MAG-D: A multivariate attention network based approach for cloud workload forecasting," Future Generation Computer Systems, vol. 142, pp. 376–392, 2023. https://doi.org/10.1016/j.future.2023.01.00

[9] I. Maiyza, N. O. Korany, K. Banawan, H. A. Hassan, and W. M. Sheta, "VTGAN: hybrid

generative adversarial networks for cloud workload prediction," Journal of Cloud Computing, vol. 12, 97, 2023. https://doi.org/10.1186/s13677-023-00473-z

[10] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, A. Basuhail, and K. Jambi, "Deep reinforcement learning for workload prediction in federated cloud environments," Sensors, vol. 23, no. 15, 6911, 2023. https://doi.org/10.3390/s23156911

[11] R. Karthikeyan and B. Sundaravadivazhagan, "COSCO2: AI-augmented evolutionary algorithm based workload prediction framework for sustainable cloud data centers," Transactions on Emerging Telecommunications Technologies, vol. 34, no. 1, e4652, 2023. https://doi.org/10.1002/ett.4652

[12] Y. Garí, D. A. Monge, E. Pacini, C. Mateos, and C. García Garino, "Reinforcement learning-based application autoscaling in the cloud: A survey," Engineering Applications of Artificial Intelligence, vol. 102, 104288, 2021. https://doi.org/10.1016/j.engappai.2021.104288

[13] S. Xue et al., "A meta reinforcement learning approach for predictive autoscaling in the cloud," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), pp. 4290–4299, 2022. https://doi.org/10.1145/3534678.3539063

[14] M. Abdullah, W. Iqbal, J. Ll. Berral, J. Polo, and D. Carrera, "Burst-aware predictive autoscaling for containerized microservices," IEEE Transactions on Services Computing, 2020. https://doi.org/10.1109/TSC.2020.2995937

[15] S. S. Mondal, N. Sheoran, and S. Mitra, "Scheduling of time-varying workloads using reinforcement learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 10, pp. 9000–9008, 2021. https://doi.org/10.1609/aaai.v35i10.17088

[16] Belgacem, S. Mahmoudi, and M. Kihl, "Intelligent multi-agent reinforcement learning model for resources allocation in cloud computing," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 2391–2404, 2022. https://doi.org/10.1016/j.jksuci.2022.03.016

[17] Karat and R. Senthilkumar, "Optimal resource allocation with deep reinforcement learning and greedy adaptive firefly algorithm in cloud computing," Concurrency and Computation: Practice and Experience, vol. 34, no. 4, e6657, 2022. https://doi.org/10.1002/cpe.6657

[18] P. K. Bal, S. K. Mohapatra, T. K. Das, K. Srinivasan, and Y.-C. Hu, "A joint resource allocation, security with efficient task scheduling in cloud computing using hybrid machine learning techniques," Sensors, vol. 22, no. 3, 1242, 2022. https://doi.org/10.3390/s22031242

[19] S. Swarup, E. M. Shakshuki, and A. Yasar, "Task scheduling in cloud using deep reinforcement learning," Procedia Computer Science, vol. 184, pp. 42–51, 2021. https://doi.org/10.1016/j.procs.2021.03.016

[20] G. Rjoub, J. Bentahar, O. Abdel Wahab, and A. Saleh Bataineh, "Deep and reinforcement learning for automated task scheduling in large-scale cloud computing systems," Concurrency and Computation: Practice and Experience, vol. 33, no. 23, e5919, 2021. https://doi.org/10.1002/cpe.5919

[21] Katal, S. Dahiya, and T. Choudhury, "Energy efficiency in cloud computing data centers: a survey on software technologies," Cluster Computing, vol. 26, no. 3, pp. 1845–1875, 2023. https://doi.org/10.1007/s10586-022-03713-0

[22] J. Díaz-de-Arcaya, A. I. Torre-Bastida, G. Zárate, R. Miñón, and A. Almeida, "A joint study of the challenges, opportunities, and roadmap of MLOps and AIOps: A systematic survey," ACM Computing Surveys, vol. 56, no. 4, Article 84, 2024. https://doi.org/10.1145/3625289

[23] Rossi, A. Visentin, D. Carraro, S. Prestwich, and K. N. Brown, "Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning," Cluster

Computing, vol. 28, Article 258, 2025. https://doi.org/10.1007/s10586-024-04933-2

[24] Ministry of Communications and Information Technology, Kingdom of Saudi Arabia, “KSA Cloud First Policy,” 2020.

[25] Saudi Vision 2030, National Transformation Program, “Annual Report 2022: Digital Transformation,” Kingdom of Saudi Arabia, 2022.