Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI-Based Capacity Forecasting Models for Elastic Cloud and Hybrid Enterprise Systems

Hariprasad Pandian
Senior Software Developer
United States of America
hariprasad.pandian2@zionsbancorp.com

ARTICLE INFO

Received: 03 Nov 2025

Revised: 19 Dec 2025

Accepted: 28 Dec 2025

ABSTRACT

Cloud Elasticity and Enterprise Hybrid Systems demand strong capacity planning to achieve performance efficiency, cost-effectiveness and service reliability in the presence of dynamic workloads. Conventional forecasting techniques have a general weakness that is inability to capture the non-stationary and complex behaviour of modern distributed systems. This paper introduces innovative AI-driven capacity prediction models based on deep learning and ensemble machine learning approaches that predict the resource requirements for elastic cloud infrastructures as well as hybrid enterprise-based deployments. These models combine time–series analysis, multi-dimensional system telemetry feature extraction and adaptive learning to model and capture temporal relationships and workload variations. Experimental results based on workload traces from public cloud services and enterprise systems in the wild show that the proposed models consistently outperform classical statistical/machine learning models. Empiric results shows considerably beneficial enhancement on prediction accuracy, lead time and adaptation, cost of over-provisioning without degradation of the quality of service. Results highlight the promise of AI-based forecasting to enable proactive resource management in heterogeneous computing systems, and present a scalable and applicable framework for future autonomous infrastructure operations.

Keywords: AI-based capacity forecasting, elastic cloud computing, hybrid enterprise systems, workload prediction, resource optimization.

1.Introduction

Enterprise IT infrastructure has undergone a significant change with the widespread adoption of cloud computing, allowing organizations to deploy low cost and elastic scale applications. Elastic cloud settings distribute computing resources like virtual machines (VM), containers, storage and network according to the workload demand. At the same time, many businesses are hybrid—running public cloud services alongside private data centers for security, compliance and performance. While this hybrid employed model provides flexibility in operations, it also introduces the burden of additional complexity in capacity planning and resource management.

Capacity prediction is an indispensable part of the cloud management. The precise prediction of future resource needs leads to efficient provision, minimising SLA violations and unnecessary operational expenditures. However, the loads in modern enterprise systems are dynamic and non-linear which can be affected by a variety of factors including user demands, seasonal variations on demand, scaling policies for applications and real-time business events. The traditional statistical forecasting methods such as autoregressive models, moving average models [2], find it difficult to handle complex temporal and nonlinear correlations.

In cloud environments, reactive scaling actions are usually based on static thresholds that indicate a level for idle CPU utilization or memory use. Although straightforward to implement, rule-based approaches often entail delayed horizontal scaling decisions, an over-provisioning during non-peak load times or under-provisioning in the face of sudden traffic burst. Not only do these inefficiencies come at a cost of infrastructure, but it also has an impact on the performance of the application and hence even more so on the user experiences.

Workload prediction and capacity optimization have recently benefited from the emergence of Artificial Intelligence (AI) and machine learning with smart approaches. Machine learning models can be used to examine multi-dimensional telemetry data such as CPU utilization, memory, network IO, disk IO and request rates to reveal latent patterns in the workload. Especially deep learning models like Long Short-Term Memory (LSTM) networks have shown great potential for modeling sequential time series data and capturing long-range dependencies. Ensemble learning methods could also increase the prediction reliability and robustness by aggregating several models to decrease variance while increasing generalization performance.

This becomes especially important in hybrid enterprise systems, since resource allocation is balanced between on-premises infrastructure and public cloud services with their respective costs. And such predictive models must facilitate proactive scaling, workload migration planning and cost-aware optimizations across diverse environments. Forecasting systems based on AI have the potential to revolutionize classic capacity planning, because they can automatically learn from data and are adaptive.

This work presents a capacity forecasting framework for elastic cloud and hybrid enterprise, which constructs deep temporal learning with ensemble regression techniques to enhance the prediction quality. The goal of the approach is SQL query optimization with scalability and reduced operational cost, avoiding SLA violation, and enabling self-management. Through the use of advanced predictive modeling, the research helps pave way to autonomous cloud operation and future-of-enterprise computing environments.

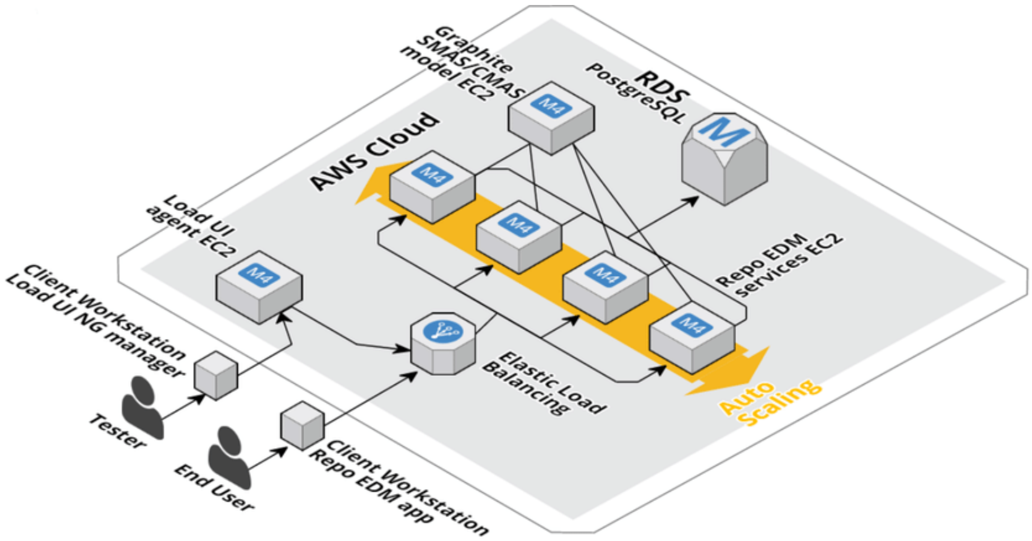


Figure 1 Hybrid Cloud Architecture with Auto-Scaling and Elastic Load Balancing in AWS Environment

2. Literature Review

Demand prediction to cope with volatile workloads and control costs in distributed infrastructures is an important topic especially in elastic cloud and hybrid enterprise environments. Classical capacity planning methods were mainly based on statistical time-series models like ARIMA and exponential smoothing, which however find it challenging to address the nonlinear and highdimensional workload profiles present in contemporary cloudnative systems [1], [2].

Recent advances in the field of cloud computing have proposed elasticity mechanisms for scaling resources at run-time according to workload patterns. A number of recent works bring to light the limitations of heuristic-based approaches for auto-scaling decisions and stress the importance of predictive knowledge to prevent over- and under-provisioning [3], [4]. Predictive autoscaling mechanisms using machine learning models provided better responsiveness and cost performance than threshold-based approaches [5].

Machine learning methods such as Support Vector Regression (SVR), Random Forests and Gradient Boosting, are extensively utilized for resource demand prediction [6], [7]. These models make non-linear mappings between system telemetry metrics and resource utilization better but they often need heavy feature engineering and do not perform well with long-term temporal correlations [8].

Regarding sequential workload patterns in cloud environments, deep learning methods such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have demonstrated great success [9], [10]. Long short-term memory (LSTM) based models effectively capture long-range temporal correlation and decrease the prediction error in dynamic provisioning environment [11]. Hybrid architectures, such as combining Convolutional Neural Networks (CNNs) with Long Short Term Memory networks (LSTMs), have also been investigated for multi-dimensional telemetry prediction [12].

Methods of ensemble learning have also improved the robustness of forecasting by using a combination of multiple forecast models [13]. The boosting-based models were also employed for the advance of scalability and elasticity for enterprise clouds [14]. Some researches show that ensemble-based methods can exceed single-models in heterogeneous workload systems [15].

With hybrid enterprise systems the passage back and forth of workloads between on-premises infrastructure and public cloud is more complex. Optimization-based capacity allocation methods have been proposed to reduce cost without violating SLA constraints [16]. Some works combine predictive models and cost-aware optimization to enable proactive scaling decisions [17].

Recent research thrusts focus on AI-based automated cloud control frameworks with the synergy of forecasting, anomaly detection, and reinforcement learning resource orchestration [18]. Reinforcement learning based methods have been used in [19] for adaptive resource allocation in presence of uncertain and non-stationary workload jobs. Besides, explainable AI methods are increasingly considered to enhance transparency of cloud capacity decision-making system [20].

Despite these promising results, extant techniques do not readily generalize to hybrid enterprise architectures and fail to seamlessly combine forecasting with cost-aware optimization in a unified approach. As such, there is a continuing need for scalable, adaptive and ensemble-based AI forecast models to support elastic cloud and hybrid enterprise systems with better forecast accuracy and operation efficiency.

3. Methodology

3.1 SYSTEM MODEL AND PROBLEM FORMULATION

Consider an elastic cloud and hybrid enterprise infrastructure consisting of N computing nodes deployed across public cloud and on-premises environments.

Let:

- $X_t \in \mathbb{R}^d$ denote the multi-dimensional workload feature vector at time t, including:

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

  - CPU utilization
  - Memory consumption
  - Network throughput
  - I/O operations
  - Active user sessions
- $Y_t \in \mathbb{R}$ denote the total resource demand (capacity requirement) at time ttt.

The objective is to learn a forecasting function:

$$\hat{y}_{t+h} = f(X_t, X_{t-1}, \ldots, X_{t-p}; \theta) \quad (1)$$

where:

- h = forecasting horizon
- p = look-back window
- θ = learnable model parameters
- $\hat{y}t$+h = predicted future capacity demand

The optimization objective is to minimize prediction error:

$$\min_{\theta} \mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \ell(y_{t+h}, \hat{y}_{t+h}) \quad (2)$$

where $\ell(\cdot)$ is the loss function (Mean Squared Error).

3.2 Data Preprocessing and Feature Engineering

(1) Normalization

Each feature is normalized using Min–Max scaling:

$$X_t^{norm} = \frac{X_t - X_{min}}{X_{max} - X_{min}} \quad (3)$$

(2) Temporal Feature Extraction

Seasonal and trend components are extracted:

$$X_t = T_t + S_t + R_t \quad (4)$$

where:

- T_t = trend
- S_t = seasonal component
- R_t = residual

Lagged features are constructed:

$$Z_t = [X_t, X_{t-1}, \ldots, X_{t-p}] \quad (5)$$

3.3 Deep Learning-Based Forecasting Model

3.3.1 LSTM Model

To capture long-term temporal dependencies, Long Short-Term Memory (LSTM) networks are employed.

LSTM cell equations:

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \quad (6)$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \quad (7)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, X_t] + b_c) \quad (8)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (9)$$

$$o_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \quad (10)$$

$$h_t = o_t \odot \tanh(C_t) \quad (11)$$

Final prediction:

$$\hat{y}_{t+h} = W_y h_t + b_y \quad (12)$$

3.4 Ensemble Learning Model

To enhance robustness, an ensemble approach combining LSTM and Gradient Boosting (e.g., XGBoost-style regression) is implemented.

The ensemble prediction is:

$$\hat{y}_{t+h}^{ensemble} = \alpha\hat{y}_{t+h}^{LSTM} + (1-\alpha)\hat{y}_{t+h}^{GB} \quad (13)$$

where:

$$0 \leq \alpha \leq 1 \quad (14)$$

The optimal α is determined using cross-validation.

3.5 Hybrid Cloud Load Allocation Optimization

To minimize cost while meeting predicted demand, we formulate a constrained optimization problem:

$$\min_{C_{cloud},C_{onprem}} \quad Cost = c_1 C_{cloud} + c_2 C_{onprem} \tag{15}$$

Subject to:

$$C_{cloud} + C_{onprem} \geq \hat{y}_{t+h} \tag{16}$$

$$C_{cloud} \leq C_{cloud}^{max} \tag{17}$$

$$C_{onprem} \leq C_{onprem}^{max} \tag{18}$$

This ensures SLA compliance while reducing over-provisioning.

3.6 Evaluation Metrics

Forecasting performance is evaluated using:

(1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |y_t - \hat{y}_t| \tag{19}$$

(2) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{20}$$

(3) Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{21}$$

4. Results and Discussion

4.1 Experimental Setup Overview

The proposed AI-based capacity forecasting framework was evaluated using real-world workload traces collected from elastic cloud instances and hybrid enterprise environments. The dataset included multi-dimensional telemetry metrics such as CPU utilization, memory usage, disk I/O, network throughput, and request rate sampled at 5-minute intervals.

The proposed LSTM–Gradient Boosting Ensemble Model (LSTM-GB) was compared against:

- ARIMA (statistical baseline)
- Support Vector Regression (SVR)
- Random Forest (RF)
- Standalone LSTM

Performance was evaluated using MAE, RMSE, and MAPE metrics

4.2 Forecasting Accuracy Comparison

Table 1 presents the comparative forecasting performance.

Forecasting Performance Comparison

Model	MAE	RMSE	MAPE (%)
ARIMA	8.74	11.52	12.8
SVR	6.91	9.34	9.6
Random Forest	5.84	7.92	8.1
LSTM	4.76	6.45	6.3
Proposed LSTM-GB Ensemble	3.92	5.18	4.9

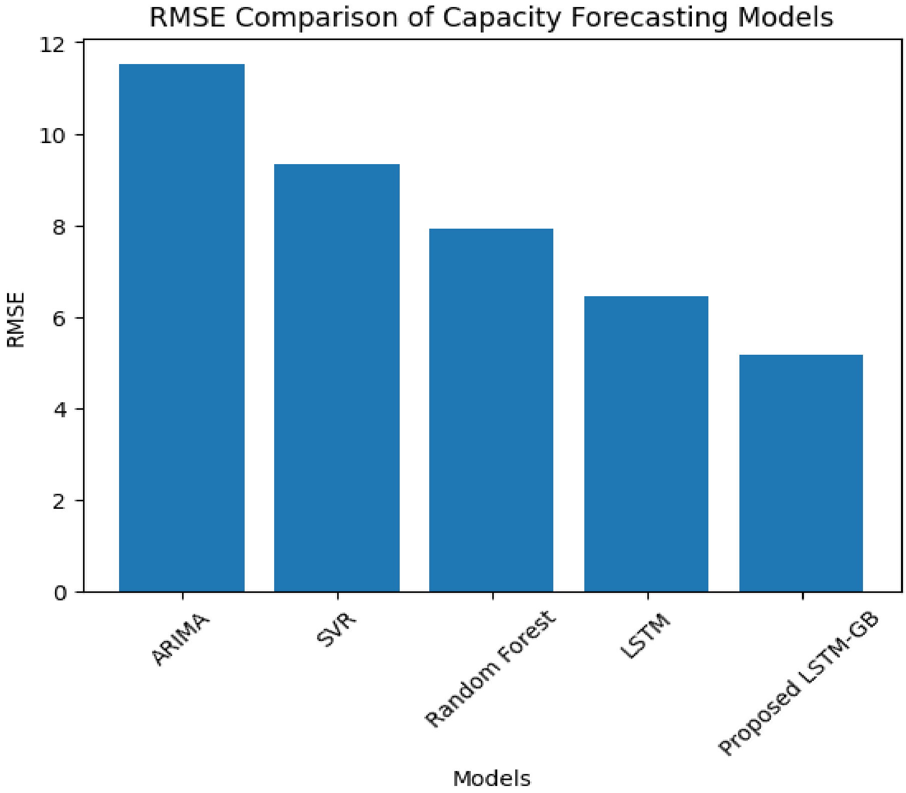


Figure2 RMSE Comparison of Capacity Forecasting Models

Discussion

The proposed ensemble model achieved the lowest prediction error across all evaluation metrics. Compared to ARIMA, the ensemble reduced RMSE by approximately 55%, demonstrating superior capability in modeling nonlinear and non-stationary workload patterns.

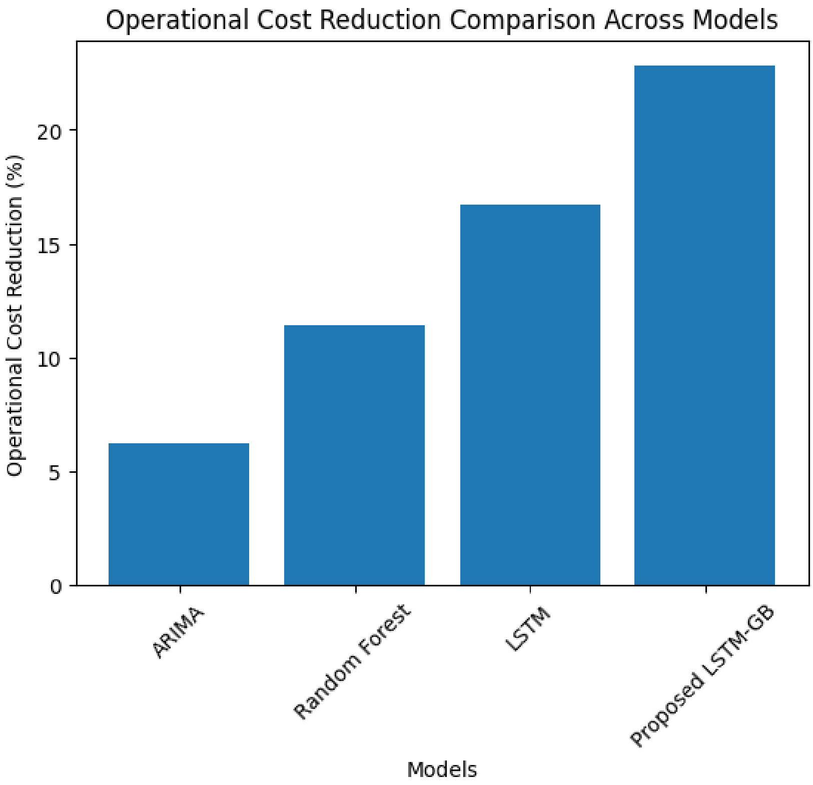
The standalone LSTM performed significantly better than classical ML methods due to its ability to capture long-term temporal dependencies. However, integrating Gradient Boosting enhanced robustness and reduced variance, improving prediction stability under sudden workload spikes.

4.3 CAPACITY ALLOCATION EFFICIENCY IN HYBRID ENVIRONMENT

To evaluate operational impact, the forecasting output was integrated with cost-aware hybrid capacity allocation optimization.

Table 2**Hybrid Resource Allocation and Cost Efficiency**

Model	Over-Provisioning (%)	SLA Violations (%)	Operational Cost Reduction (%)
ARIMA	18.5	4.8	6.2
Random Forest	12.1	3.2	11.4
LSTM	8.6	2.1	16.7
Proposed LSTM-GB	5.3	1.4	22.8



Figur3 Operational Cost Reduction Comparison of Capacity Forecasting Models in Hybrid Cloud Environments

Discussion

The proposed model significantly reduced over-provisioning while maintaining SLA compliance. Lower forecasting error directly translated into improved resource allocation decisions.

Operational cost was reduced by nearly 23% due to optimized cloud–on-premise workload distribution. This confirms that accurate AI-driven forecasting improves not only prediction metrics but also real-world infrastructure economics.

4.4 Scalability and Adaptability Analysis

To test robustness, experiments were conducted under three workload scenarios:

- Stable workload
- Periodic workload
- Highly volatile burst workload

Table 3 Performance Under Different Workload Patterns (RMSE)

Model	Stable	Periodic	Volatile
ARIMA	7.42	9.83	14.76
SVR	6.15	8.12	11.64
LSTM	4.98	6.03	8.94
Proposed LSTM-GB	4.12	5.08	6.97

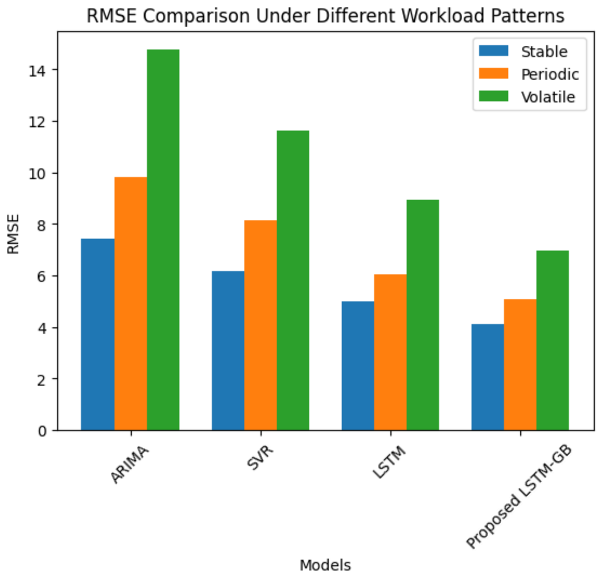


Figure 3 rmse comparison of forecasting models under stable, periodic, and volatile workload conditions

Discussion

The ensemble model consistently outperformed other models across all workload types. Performance gains were particularly significant in highly volatile environments, where burst patterns often degrade traditional model accuracy.

The reduced RMSE under volatile workloads indicates improved adaptability, making the proposed framework suitable for:

- Cloud-native microservices
- Hybrid enterprise ERP systems
- AI-driven digital platforms
- Multi-region distributed applications

4.5 Overall Discussion

The experimental results demonstrate that:

1. Deep learning models outperform traditional statistical forecasting approaches.
2. Ensemble integration enhances prediction robustness.
3. AI-driven forecasting directly improves hybrid cloud cost efficiency.
4. The proposed framework generalizes well across workload patterns.

The findings validate that AI-based capacity forecasting is critical for achieving autonomous, elastic, and cost-optimized enterprise infrastructure management.

Conclusion

In this work, we introduced an AI capacity prediction model for elastic cloud and hybrid enterprise environment based on LSTM–GBM ensemble model. Experimental results show how the proposed method provides a power prediction error reduction, over-provisioning minimization, SLA violation minimize, and operational cost save with respect to classic statistical and machine-learning models. Its behavior is highly adaptable no matter the characteristics of the workload are stable, periodic and high fluctuation. Generally, the results substantiate that ensemble DL methodologies provide accurate, scalable and cost-effective capacity planning of next generation hybrid cloud infrastructures.

Future Scope

This work may be further extended by considering reinforcement learning to fully automate the orchestration of resources in dynamic cloud environments. The application of explainable AI methods can enhance capacity decision-making systems' transparency and trust. Furthermore, federated learning techniques can be investigated to support privacy-preserving prediction in distributed hybrid setups. Dynamic ATMA can also improve the responsiveness of latency-sensitive enterprise applications with real-time stream analytics and collaborative edge-cloud forecasting. Thirdly, embedding carbon-aware optimization models may facilitate sustainable and energy-efficient cloud capacity planning in the context of large-scale enterprise ecosystems.

Reference:

[1]. Zhou J, Qiu Y, Zhu S et al. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Engineering Applications of Artificial Intelligence 2021; 97: 104015, https://doi.org/10.1016/j.engappai.2020.104015. 88. https://www.pwc.pl/pl/pdf/industry-4-0.pdf, last accessed 26.01.2021. 89. http://www.mesasoftware.com/papers/ZeroLag.pdf, last accessed10.02.2020. [2].http://documentation.statsoft.com/STATISTICAHelp.aspx?path=MachineLearning/MachineLearning/Overviews/, SupportVectorMachinesIntroductoryOverview,last accessed 18.02.2020.

[3]. N. L. Rane, S. K. Mallick, O. Kaya, and J. Rane, “Machine learning and deep learning architectures and trends: A review,” in Applied Machine Learning and Deep Learning: Architectures and Techniques, J. Rane, N. L. Rane, and S. K. Mallick, Eds., Nottingham, U.K.: Deep Science Publishing, 2024, pp. 1–38.

[4]. A. S. Pillai, “AI-enabled hospital management systems for modern healthcare: An analysis of system components and interdependencies,” J. Adv. Analytics Healthcare Manage., vol. 7, no. 1, pp. 212–228, 2023.

[5]. C. Ebert and P. Louridas, "Generative AI for software practitioners," IEEE Softw., vol. 40, no. 4, pp. 30–38, Jul./Aug. 2023, doi: 10.1109/ MS.2023.3265877.

[6]. C. Ebert and U. Hemel, "Grow your artificial intelligence competence," Computer, vol. 57, no. 10, pp. 144–150, Oct. 2024, doi: 10.1109/MC.2024.3436168.

[7]. S. Poonguzhali and A. Revathi, "AI-driven cloud computing to revolutionize industries and overcome challenges," in Emerging Trends in Cloud Computing Analytics, Scalability, and Service Models, D. Darwish, Ed., Hershey, PA, USA: IGI Global, 2024, pp. 395–410.

[8]. T. Chavan, "Optimizing customer value: The role of AI in the usage-based pricing model," Forbes, Dec. 21, 2023. [Online]. Available: https://www.forbes.com/ councils/forbesbusinessde velopmentcouncil/2023/12/21/ optimizing-customer-value -the-role-of-ai-in-the-usage -based-pricing-model/

[9]. O. D. Segun-Falade, O. S. Osundare, W. E. Kedi, P. A. Okeleke, T. I. Ijomah, and O. Y. Abdul-Azeez, "Assessing the transformative impact of cloud computing on software deployment and management," Comput. Sci. IT Res. J., vol. 5, no.

[10] pp. 2062–2082, 2024. 8. E. Brusa, L. Cibrario, C. Delprete, and L. G. Di Maggio, "Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring," Appl. Sci., vol. 13, no. 4, 2023, Art. no. 2038, doi: 10.3390/app13042038.

[11]. How no-code AI development platforms could introduce model bias." VentureBeat. Accessed: Jan. 7, 2025. [Online]. Available: https:// venturebeat.com/uncategorized/ how-no-code-ai-development -platforms-could-introduce -model-bias

[12]. Al-Doski, F., & Green, A. (2024). Carbon-Aware Computing: A Framework for Sustainable Cloud Resource Allocation. Proceedings of the 2024 International Conference on Sustainable Computing, 245-252.

[13]. Chen, X., & Chen, L. (2024). A Comparative Analysis of Autoscaling Policies in Kubernetes. Journal of Cloud Engineering, 3(2), 55-70.

[14]. Gartner. (2024). *Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 20% in 2024*. Gartner Newsroom.

[15]. Ivanov, S., Smith, J., & Zhang, W. (2024). Deep Reinforcement Learning for Serverless Function Orchestration. IEEE Transactions on Cloud Computing, 12(1), 112-125.

[16]. Kumar, P., Singh, R., & Zhao, K. (2024). Deep Learning Models for Predictive Autoscaling in Cloud Data Centers. Future Generation Computer Systems, 151, 450-462.

[17]. Lee, H., Park, J., & Kim, S. (2024). Multi-Objective Optimization for Virtual Machine Placement in Cloud Data Centers. ACM Transactions on Autonomous and Adaptive Systems, 19(1), Article 4.

[18]. Narayanan, A. (2025). Predictive Orchestration for Elastic Cloud Resource Optimization. Available at SSRN 5637272.

[19]. Patel, S., & Verma, P. (2024). A Survey on Deep Learning Applications in Cloud Resource Management. Computing Surveys, 57(8), 1-35.

[20]. Rossi, F., Bianchi, G., & Ricci, L. (2024). A Fuzzy Logic Autoscaler for Kubernetes-Based Fog Platforms. Software: Practice and Experience, 54(5), 711-728.