International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., S1, 2026
January

Special issue

ISSN: 2321-7758

Adaptive Cloud Resource Management in Dynamic Environments Using AI Techniques

Pusapati Viswa Jyothi[1], Penta Vijaya Raghava Kumari Mounika[2], Somana Sunitha[3], Nallamilli Kaavya Tejaswi[4]

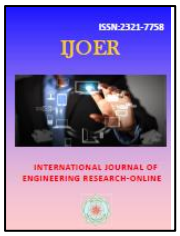
[1] Lecturer in Computer Applications, Pithapur Raja's Government College (Autonomous), Kakinada, Andhra Pradesh, India

[2] Lecturer in Artificial Intelligence and Machine Learning, Srinivasa Institute of Engineering and Technology, Andhra Pradesh, India

[3] Lecturer in Computer Science and Engineering, Srinivasa Institute of Engineering and Technology, Andhra Pradesh, India

[4] M.Tech in Computer and Communication Engineering, Jawaharlal Nehru Technological University College of Engineering, Kakinada, Andhra Pradesh, India

DOI: 10.33329/ijoer.14.S1.146



Abstract

Efficient cloud resource management is a critical challenge due to random workloads and heterogeneous infrastructures, which are frequently leading to resource over-provisioning or service-level agreement (SLA) violations. Reactive methods often cause over- or under-provisioning, thus inflating the costs. This paper analyses deep learning models like CNN, LSTM, and Transformers for estimates resource utilities and borrowing Google Cluster Data (GCD) to assess predictive accuracy and adaptability. Multiple experiments conclude that Transformers give better results while capturing complex temporal patterns, outperforming CNN and LSTM in metrics like MAE, RMSE, and resource utilization (up to 88%). The Transformer model enables precise scaling, minimizing bottlenecks and optimizing the costs. These findings emphasise on using the advanced neural architectures for intelligent cloud management, empowering providers to enhance utilization, reduce expenses, and boost reliability amid rapid digital evolution.

Keywords: Resource Management, Cloud computing, Machine learning, Deep learning, Convolutional Neural Networks (CNN).

1.Introduction

The speedy enhancements and techniques in the field of cloud computing and large-scale distributed systems, it has become very difficult to manage the complex computational resources with good efficiency. In order to mitigate the issues such as fluctuating workloads, performance variability, and cost optimization challenges, the Predictive resource allocation has emerged as a critical solution. We also need to leverage the Deep learning techniques which has gained significant attention in this domain due to their ability to model complex, nonlinear patterns from large volumes of historical and real-time data. Traditional statistical and machine learning

algorithms are highly accurate likewise deep learning models can automatically extract high-level features and capture time dependencies in workload behaviour.[1] Techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Transformer models have been widely implemented to evaluate and forecast resource demands including CPU usage, memory consumption, and network bandwidth. Deep learning techniques would widely help us forecasting t and accurately predicting future workloads, thus enable proactive and adaptive resource provisioning, improving utilization efficiency, reducing operational costs, and ensuring compliance with service-level agreements (SLAs)[1][2]. Due to drastic and continuous improvements in cloud environments and its evolution, deep learning plays a very key role in enabling intelligent, scalable, and independent resource management systems. Achieving optimal efficiency, cost-effectiveness, and better performance in cloud computing environments remains difficult despite advancements in predictive resource allocation.

2. Literature Review

The literature review presents a consolidated summary of existing research that gives the understanding of resource allocation mechanisms in cloud computing environments. It emphasis the critical role of reliable resource management and explores a variety of approaches, frameworks, and models proposed to address the challenges associated with dynamic and complex cloud infrastructures.

S. R. Swain et al. [9] emphasis on the effective allocation and utilization of resources in cloud environments. The authors put forth different resource allocation strategies and aim to maximize resource utilization with low cost but efficient and enable high-performance cloud services. The study focuses on the details of managing cloud resources.

S. A. Murad et al. [10] worked on analysing the various job scheduling methods in cloud computing.

It showcases a highly advanced framework that utilizes priority rules to schedule jobs and allocate resources. The research explores the current hurdles of job scheduling in cloud environments and highlights the dire need of intelligent frameworks in optimizing resource allocation.

3. Methodology

The proposed approach for the predictive resource allocation in cloud computing implements a well-structured and systematic workflow which is designed to enable accurate workload forecasting and efficient resource management. To assess the efficiency of DL algorithms like CNN, LSTM, and Transformer model for resource allocation in cloud computing environments efficiently. The objective is to learn more about the algorithm that shows the best predictive accuracy and flexibility even in the hostile workload conditions. This projects predictive resource allocation strategies, using DL algorithms using Google Cluster Data (GCD) dataset. The research evaluates the effectiveness of these algorithms for improving resource allocation in cloud computing environments. [1,2] This technique ensures proactive, scalable and cost-effective cloud resource management.

4. System Architecture

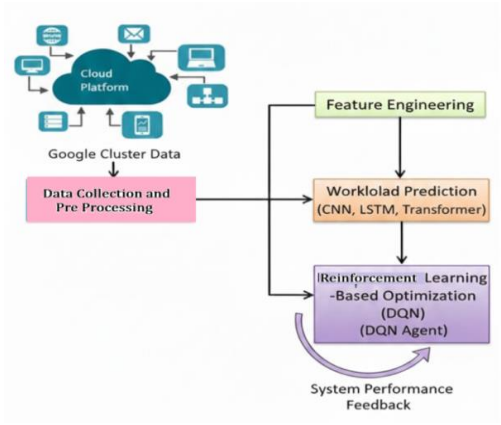


Figure1. System Architecture

The above Figure1 gives a predictive resource allocation framework for cloud computing using ML algorithms, showing the entire workflow from data accretion to intelligent optimization with feedback. The process starts with Google Cluster Data obtained from the cloud platform, representing real-world workload traces. This data is fed as input to the Data Collection and Pre-processing module, where the original workload information is wiped, normalized, and trained for analysis.

As part of the next phase, the Feature Engineering extracts meaningful time-lined and statistical characteristics from the pre-processed data. These engineered and combined features are then fed into the Workload Prediction module, which then employs deep learning models such as CNN, LSTM, and Transformer models to predict the future resource allocations. The forecasted workloads are subsequently given to the Reinforcement Learning–based Optimization component, and thus implemented using a Deep Q-Network (DQN) agent. This DQN agent summerizes the intelligent resource allocation and scaling decisions based on the forecasted demand and current system state.

At last, the System Performance Feedback loop plays a vital role to approach the real-time performance metrics back to the DQN agent, which attains a constant learning and adaptive optimization. This closed-loop architecture makes sures a systematic, proactive, and SLA-aware cloud resource management system by continuously refining forecasting and allocation strategies.

5.Evaluation Parameters

The complete performance of this predictive resource allocation frameworks is mainly hanged on deep learning and is estimated leveraging a combination of prediction accuracy and resource efficiency. These metrics ensure an accurate assessment of both forecasting capability and operational impact.[2,3]Prediction Accuracy Metrics are used to compare how accurately the predicted resource demand matches actual usage. Common measures include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). Lower error values and higher R^2 scores implicate the better predictive performance.

Resource Utilization Metrics gives us the idea on how effectively the cloud resources are used. These include different parameters like average CPU utilization, memory utilization, and network bandwidth usage. Improved utilization also reflects the reduced over-provisioning and under-utilization. Together, these evaluation parameters provide a holistic view of the effectiveness, robustness, and Practicality of deep learning–based predictive resource allocation in cloud environments. In this paper we are focussing only on prediction accuracy and resource utilization.

Mean Absolute Error (MAE): MAE is a direct measurement that computes the average absolute deviation between the predicted and actual values. MAE measure precisely and conclusively to prediction model performs where lower MAE values represents higher model performance. Eq.1 represents MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad \ldots 1$$

Where n= "data points", $\hat{y}i$= "predicted values", yi= "actual values"

Root Mean Squared Error (RMSE):

RMSE calculates the square root of the mean of the squared differences between predicted and actual values.RMSE is more accurate at predicting errors when compared to MAE.This makes it a most optimal tool for detecting outliers or significant discrepancies. Eq.2 represents RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad \ldots 2$$

Where n= "data points", $\hat{y}i$= "predicted values", yi= "actual values".

Normalized Mean Absolute Error (NMAE): NMAE measures the average absolute difference between predicted and actual values which is normalized to a specific range. It is derived from the MAE metric. The NMAE provides a substantial measure of the accuracy of predictions, making it appropriate for comparing models across various. Eq.3 represents NMAE.

$$NMAE = \frac{MAE}{\max(y) - \max(x)} \quad \ldots 3$$

Normalized Root Mean Squared Error (NRMSE):

NRMSE is a more normalized version of the Root Mean Squared Error (RMSE), almost similar to the NMAE. The method divides the RMSE by the range of actual values, resulting in a relative measure of prediction accuracy that can be compared between different datasets.Eq.4 represents NRMSE.

$$NRMSE = \frac{RMSE}{\max(y) - \max(x)} \quad \ldots 4$$

Resource Utilization (%):

Resource Utilization is a metric which tells about the effectiveness of resource allocation. It calculates the proportion of resources that were effectively utilized from the overall allocated resources. Thereby, High resource utilization indicates effective allocation, whereas low utilization might suggest inefficiency and wastage. Eq.5 represents Resource utilization.

$$Resource\ Utilization = \frac{Used\ resources}{Allocated\ Resources} \times 100\% \quad .$$

6. RESULTS

Table 1. Prediction Accuracy Metrics

Model	MAE	RMSE	NMAE	NRMSE
CNN	0.07	0.25	0.2	0.18
LSTM	0.04	0.17	0.08	0.2
Tranformer	0.02	0.1	0.05	0.09

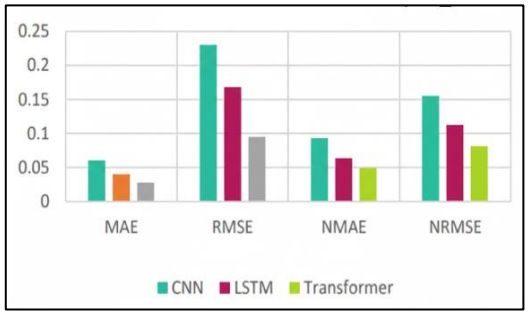


Figure. Prediction Accuracy Metrics

Table 2. Resource Utilization

Model	Resource Utilization (%)
CNN	87
LSTM	94
Tranformer	96

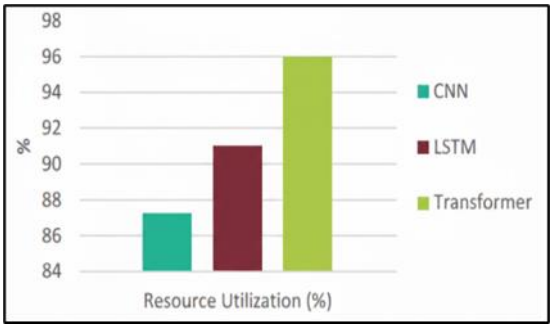


Figure. Resource Utilization

Transformer in cloud computing resource allocation prediction are shown in table-1,2 and figure 2,3. CNN makes good predictions with an MAE of 0.07 and an RMSE of 0.25.The NMAE and NRMSE scores of 0.2 and 0.18 indicate high relative accuracy. Resource Utilization is 87%, indicating efficient resource allocation. The LSTM model has a 0.04 MAE and 0.17 RMSE, indicating a slight prediction accuracy improvement. The NMAE and NRMSE values of 0.08 and 0.2 indicate a higher level of relative accuracy. This shows a remarkable 94% accuracy in narrow downing the targets. The 93% resource utilization rate implies a good resource allocation. With an MAE of 0.02 and RMSE of 0.1, the Transformer algorithm is more accurate.

High relative accuracy is noticed by observing by the NMAE and NRMSE values of 0.05 and 0.09.High Resource Utilization of 96% indicates very efficient resource allocation. Thus, Transformer outperforms LSTM in predictive resource allocation accuracy and efficiency. CNN is adequate, but the Transformer model is better for cloud computing resource allocation. The Transformer model has limited prediction errors and enhanced high resource utilization.

7.Conclusion

The relative evaluation of deep learning algorithms CNN, LSTM, and Transformer for predictive resource allocation in cloud computing environments gives us an important comprehension into their relative strengths and effectiveness. Among the evaluated models, the Transformer architecture describe uperior performance, achieving the lowest Mean Absolute Error (MAE), which indicates its strong capability in accurately forecasting future resource demands. Even the LSTM model performs efficiently, showing a reliable forecast accuracy and enhanced resource utilization due to its ability to capture temporal dependencies in workload patterns. While the CNN model delivers substantial results, its performance is comparatively limited while handling long-term dependencies that might have crept in dynamic cloud workloads.[1][2]

Overall, the Transformer model emerges as the best and most reliable approach for enhanced predictive resource allocation, setting a bright benchmark for the advanced prediction techniques in cloud environments. These results emphasize the role of sophisticated deep learning models in improving reliability, efficiency and intelligent decision-making in cloud resource management. Future research can be elevated to different directions which include the development of hybrid and adaptive models, incorporating real-time learning mechanisms, more focused on energy-aware resource allocation, and extensive predictive strategies to edge and fog computing environments. Addressing different levels of security and privacy impediments and leveraging Automatic Machine Learning techniques for fully automated resource provisioning remain critical areas for further advancement.

8. Future scope

Quantum Computing adoption with this predictive resource allocation would open up a promising future direction for cloud computing research. Quantum computing can be leveraged to accelerate the potential capabilities to solve complex optimization and mitigate prediction problems substantially and hasten the classical computing due to the quantum parallelism and superposition techniques. Quantum machine learning models can further enhance prediction accuracy by processing muti-dimensional cloud workload data more efficiently. Further, quantum-inspired optimization techniques may enable near-optimal resource allocation under strict service-level agreement (SLA) constraints while mitigating energy consumption and operational costs. As hybrid quantum–classical cloud architectures emerge, predictive resource allocation frameworks can incorporate quantum accelerators to support real-time, large-scale decision-making.

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, 2022.

[2]. K. Kumaran and E. Sasikala, "Computational access point selection based on resource allocation optimization to reduce the edge computing latency," *Measurement: Sensors*, vol. 24, p. 100444, 2022.

[3]. P. Pradhan, P. K. Behera, and B. N. B. Ray, "Modified round robin algorithm for resource allocation in cloud computing," *Procedia Computer Science*, vol. 85, pp. 878–890, 2016.

[4]. Z. Sharif, L. T. Jung, M. Ayaz, M. Yahya, and S. Pitafi, "Priority-based task scheduling and resource allocation in edge

computing for health monitoring system," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 2, pp. 544–559, 2023, doi: 10.1016/j.jksuci.2023.01.001.

[5]. P. Wei, Y. Zeng, B. Yan, J. Zhou, and E. Nikougoftar, "VMP-A3C: Virtual machine placement in cloud computing based on asynchronous advantage actor–critic algorithm," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 5, p. 101549, 2023, doi: 10.1016/j.jksuci.2023.04.002.

[6]. X. Xiao, M. Zhao, and Y. Zhu, "Multi-stage resource-aware congestion control algorithm in edge computing environment," *Energy Reports*, vol. 8, pp. 6321–6331, 2022, doi: 10.1016/j.egyr.2022.04.078.

[7]. V. Khetani, Y. Gandhi, S. Bhattacharya, S. N. Ajani, and S. Limkar, "Cross-domain analysis of machine learning and deep learning: Evaluating their impact in diverse domains," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, pp. 253–262, 2023.

[8]. A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource allocation in multi-access edge computing for 5G-and-beyond networks," *Computer Networks*, vol. 227, p. 109720, 2023.

[9]. T. Thein, M. M. Myo, S. Parvin, and A. Gawanmeh, "Reinforcement learning-based methodology for energy-efficient resource allocation in cloud data centers," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, no. 10, pp. 1127–1139, 2020, doi: 10.1016/j.jksuci.2018.11.005.

[10]. Y. Xie, C. Allen, and M. Ali, "Critical success factor-based resource allocation in ERP implementation: A nonlinear programming model," *Heliyon*, vol. 8, no. 8, p. e10044, 2022.