A Review on Estimating Cloud Performance Metrics Using Machine Learning And Deep Learning Models

Surbhi Jhariya[1], Prof. Pawan Panchole[2]
[1, 2] Dept of CSE
[1, 2]VITM, Indore, India

Abstract- *Data driven cloud computing model have resulted in unprecedented paradigm shifts in cloud application development. Many applications have found data driven cloud computing models indispensable due to the need for high performance computing. Performance prediction is essential for both cloud service providers and users. Providers rely on accurate predictions to manage resources effectively, prevent over-provisioning or under-provisioning, and maintain service-level agreements (SLAs). Users, on the other hand, benefit from performance prediction when selecting cloud services that meet their application requirements. Inadequate performance prediction can lead to increased operational costs, degraded service quality, and customer dissatisfaction. Thus, robust prediction mechanisms are indispensable in ensuring the efficient operation of cloud systems. This work presents a regression learning based model for performance prediction in cloud environments. This paper presents a review on the contemporary Machine Learning and Deep Learning Models for Estimating Cloud Performance Metrics.*

Keywords: Cloud Computing, service-level agreements (SLAs). Regression Learning, Neural Network, MAPE.

I. INTRODUCTION

Cloud computing has emerged as a transformative paradigm in modern information technology, enabling organizations to access computing resources over the internet without owning or managing physical infrastructure. Instead of deploying local servers or storage systems, users can leverage remote datacenters operated by cloud service providers [1]. This shift offers significant advantages, including cost savings, rapid scalability, global accessibility, and reduced management overhead. Cloud computing fundamentally changes how individuals, enterprises, and governments build, deploy, and maintain digital services, making it a key enabler of digital transformation [2].

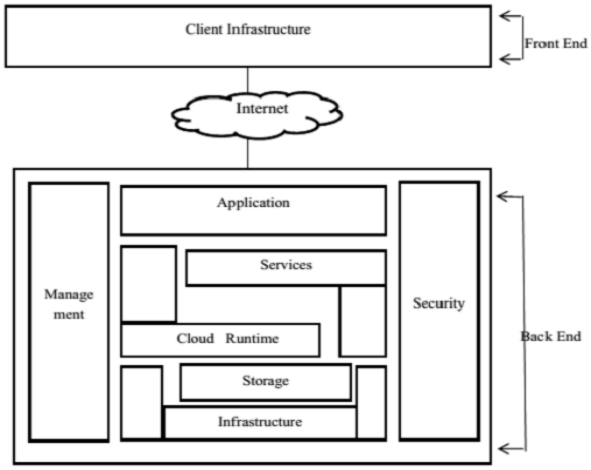

Fig.1. The Cloud Computing Architecture

Figure 1 presents the cloud computing architecture. One of the core motivations behind cloud computing is its ability to provide on-demand resources with minimal management effort. Users can provision computing power, storage, or applications in minutes rather than weeks, allowing businesses to innovate faster and respond to market dynamics with agility [3].

Cloud infrastructures are also designed with built-in redundancy and failover capabilities, enhancing service reliability and minimizing downtime. The pay-as-you-go billing approach further allows organizations to optimize costs by paying only for the resources they consume [4].

The foundation of cloud computing lies in its service delivery models, which categorize how resources are packaged and delivered. The three primary cloud service models—Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)—offer varying degrees of control, flexibility, and management responsibility. This layered model ensures that different users, from developers to business end-users, can select the appropriate level of abstraction and convenience to meet their specific needs [5]

II. MATERIALS AND METHODS

The need for the presents study stems from the fact that different cloud services are needed by different clients. Typically, cloud services are categorized as:

2.1 Cloud Services

Infrastructure as a Service (IaaS) is the most fundamental level, offering virtualized computing resources such as virtual machines, storage, and networking [6]. It gives users maximum control over the operating system, runtime environment, and applications while offloading the hardware management to the cloud provider. IaaS is suitable for organizations that require customizable infrastructure or wish to migrate legacy applications to the cloud. Popular examples include Amazon EC2, Google Compute Engine, and Microsoft Azure Virtual Machines [7]

Platform as a Service (PaaS) offers a higher level of abstraction by providing a managed environment for application development, deployment, and scaling. Developers can focus on writing code while the provider handles infrastructure provisioning, load balancing, and runtime management. PaaS is ideal for rapid application development and DevOps workflows, helping teams reduce complexity and accelerate time-to-market. Services like Google App Engine, AWS Elastic Beanstalk, and Azure App Service exemplify this model [8].

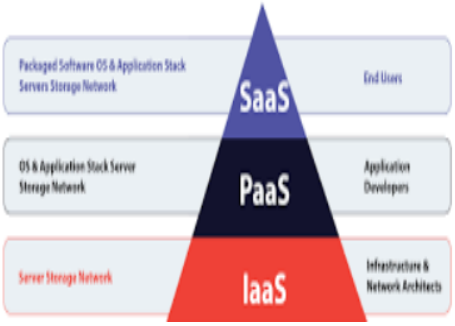

Fig.2. Different Cloud Services

Figure 2 presents the different cloud services. In addition to service models, cloud computing also defines deployment models such as public, private, hybrid, and multi-cloud. Public clouds offer shared infrastructure accessible to multiple users, while private clouds provide dedicated environments for organizations requiring strict security and customization.

Software as a Service (SaaS) delivers fully managed applications accessible through web browsers or mobile apps. Users do not need to install or maintain software; instead, updates, security, and data management are overseen by the provider. SaaS is widely used in business productivity, customer relationship management, and collaboration tools. Common examples include Google Workspace, Microsoft 365, Salesforce, and Dropbox. This model significantly reduces operational effort and ensures users always have access to the latest features [9].

Hybrid cloud combines both models, allowing data and applications to move seamlessly between environments, whereas multi-cloud leverages multiple providers to avoid vendor lock-in and enhance resilience [10].

2.2 Cloud Performance

Cloud environments are inherently dynamic, with fluctuating workloads, shared resource usage, and complex virtualization layers. These uncertainties make it difficult to guarantee predictable performance, thereby creating a strong need for accurate cloud performance prediction mechanisms. At the core of this need is the variability of cloud resources. In cloud infrastructures, multiple tenants share CPU cores, memory, storage, and network bandwidth. This multi-tenancy often leads to performance interference, where one user's workload affects another's performance. Moreover, providers deploy load balancing and auto-scaling algorithms that constantly reallocate resources based on demand. Without performance prediction, users struggle to understand how their applications will behave under different resource allocations, leading to suboptimal decisions during deployment [11].

Cloud performance prediction is also essential for cost optimization. Since most cloud services operate on a pay-as-you-go model, users must carefully select the right combination of virtual machine types, storage classes, and network configurations. Over-provisioning resources can lead to unnecessary spending, while under-provisioning may cause degraded application performance or service outages [12]. Predictive models help estimate the performance-to-cost ratio of various configurations, enabling organizations to choose the most economical options without compromising quality of service (QoS). Another important driver is the need to guarantee Service Level Agreements (SLAs). Businesses that deliver services on the cloud must meet specific latency, throughput, and availability requirements defined by SLAs. Violations can lead to financial penalties, reputational damage, and user dissatisfaction. By predicting performance ahead of time, organizations can proactively detect potential bottlenecks, allocate sufficient resources, and maintain compliance with SLAs even under peak workloads [13]

Cloud performance prediction is also crucial for scaling applications efficiently. Auto-scaling mechanisms depend on accurate forecasts of future workload demands. If the scaling decision is delayed or inaccurate, it may result in performance drops, increased response times, or sudden spikes in resource usage. Predictive models based on machine learning or statistical analysis allow systems to anticipate workload patterns, enabling proactive scaling and more efficient resource management across distributed cloud environments [14]

III. EMPLOYING DEEP LEARNING MODELS FOR PERFORMANCE ESTIMATION

Different deep learning architectures are employed depending on the nature of the performance metric and data characteristics. Feedforward deep neural networks (DNNs) are commonly used for static or short-term performance estimation, where current resource states are mapped to expected performance outcomes. Convolutional neural networks (CNNs) can extract spatial correlations among multiple virtual machines or containers, especially in large-scale data centers. Recurrent neural networks (RNNs), particularly long short-term memory (LSTM) and gated recurrent unit (GRU) models, are widely used for capturing temporal dependencies in workload patterns and predicting future performance metrics such as latency and throughput [7].

One of the key advantages of using deep learning for cloud performance estimation is its ability to adapt to workload variability. Cloud workloads often exhibit bursty, seasonal, or unpredictable behavior. Deep learning models trained on historical and real-time data can generalize across diverse workload patterns and provide accurate predictions even under sudden workload changes. This capability is especially important for proactive resource management, where performance estimation is used to trigger auto-scaling, load balancing, or migration decisions before service-level agreements (SLAs) are violated [8].

Deep learning-based performance estimation also supports intelligent cloud optimization and decision-making [9]. Accurate predictions of response time or resource utilization enable cloud providers to optimize virtual machine placement, reduce energy consumption, and improve overall system efficiency. For cloud users, performance-aware scheduling and cost-performance trade-off analysis become feasible when reliable performance estimates are available. Moreover, integrating deep learning models with reinforcement learning frameworks allows cloud systems to continuously learn optimal control policies based on predicted performance outcomes [10].

Despite its advantages, estimating cloud performance using deep learning also faces several challenges. Training deep models requires large volumes of high-quality data, which may not always be available due to privacy, security, or monitoring overhead constraints [11]. Model interpretability is another concern, as deep learning models often act as black boxes, making it difficult for cloud operators to understand the reasons behind specific predictions. Additionally, model retraining and deployment in rapidly changing cloud environments introduce computational and operational overheads [12].

IV. EXISITING WORK

This section presents a review of the previous work in the domain.

Huang et al. [14] proposed a novel machine learning-based performance prediction approach for applications running in the cloud. To achieve high-accuracy predictions for black-box VMs, the proposed method first identifies the running application inside the virtual machine. It then selects highly correlated runtime metrics as the input of the machine learning approach to accurately predict the performance level of the cloud application. Experimental results with state-of-the-art cloud benchmarks demonstrate that our proposed method outperforms existing prediction methods by more than 2× in terms of the worst prediction error. In addition, we successfully tackle the challenge of performance prediction for applications with variable workloads by introducing the performance degradation index, which other comparison methods fail to consider. The workflow versatility of the proposed approach has been verified with different modern servers and VM configurations.

Yuan et al. [15] proposed a novel prediction approach named VSBG that seamlessly and innovatively combines variational mode decomposition (VMD), Savitzky Golay (SG) filter, bi-directional long short-term memory (LSTM), and grid LSTM to predict workload and resource usage in CDCs accurately. VSBG innovatively integrates VMD and an SG filter in a four-step manner before performing its prediction. VSBG leverages VMD to divide nonstationary workload and resource time series into multiple mode functions. Then, in VSBG, this work designs a quadratic penalty, minimizes it with a Lagrangian multiplier, and adopts a logarithmic operation and the SG filter to smooth the first mode function to eliminate noise interference. Extensive experiments with different real-world data sets prove that VSBG outperforms a holistic set of state-of-the-art algorithms on prediction accuracy and convergence speed.

Yazdanian et al. [16] proposed a hybrid E2LG algorithm, which decomposes the cloud workload time-series into its constituent components in different frequency bands using empirical mode decomposition method which reduces the complexity and nonlinearity of prediction model in each frequency band. Also, a new state-of-the-art ensemble GAN/LSTM deep learning architecture is proposed to predict each sub band workload time-series individually, based on its degree of complexity and volatility. The ensemble GAN/LSTM architecture, which employs stacked LSTM blocks as its generator and 1D ConvNets as discriminator, can exploit the long-term nonlinear dependencies of cloud workload time-series effectively specially in high-frequency, noise-like components.

Jeddi et al. [17] proposed a hybrid wavelet time series decomposer and GMDH-ELM ensemble method named Wavelet-GMDH-ELM (WGE) for NFV workload forecasting which predicts and ensembles workload in different time-frequency scales. We evaluate the WGE model with three real cloud workload traces to verify its prediction accuracy and compare it with state of the art methods. The results show the proposed method provides better average prediction accuracy. Especially it improves Mean Absolute Percentage Error (MAPE) at least 8% compared to the rival forecasting methods such as support vector regression (SVR) and Long short term memory (LSTM).

Gao et al. [18] showed that meeting QoS with cost-effective resource is a challenging problem for CSPs because the workloads of Virtual Machines (VMs) experience variation over time. It is highly necessary to provide an accurate VMs workload prediction method for resource provisioning to efficiently manage cloud resources. In this paper, authors first compare the performance of representative state-of-the-art workload prediction methods. We suggest a method to conduct the prediction a certain time before the predicted time point in order to allow sufficient time for task scheduling based on predicted workload. To further improve the prediction accuracy, authors introduce a clustering based workload prediction method, which first clusters all the tasks into several categories and then trains a prediction model for each category respectively.

Chen et al. [19] proposed a deep Learning based Prediction Algorithm for cloud Workloads (L-PAW). First, a top-sparse auto-encoder (TSA) is designed to effectively extract the essential representations of workloads from the original high-dimensional workload data. Next, authors integrate TSA and gated recurrent unit (GRU) block into RNN to achieve the adaptive and accurate prediction for highly-variable workloads.. Moreover, the performance results show that the L-PAW achieves superior prediction accuracy compared to the classic RNN-based and other workload prediction methods for high-dimensional and highly-variable real-world cloud workloads.

Wang et al. [20] provided Adaptive Dispatching of Tasks in Cloud. Cloud computing domain has been witnessing a large traffic and users dependent on it. With most of the work being shifted to the internet platform, the cloud services have become dominant in all aspects of business and technology. In this work, the authors proposed a novel study of cloud tasks dispatching. There are allocation schemed and algorithms that have been used as a part of the model. The response time that is computed has been reduced considerably in this work. Various hosts have been deployed for the proper client and server interaction. The time delays were greatly lessened and it proved to be a really useful methodology.

Duggan et al. [21] presented Research on Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks. This study aims to predict the CPU consumption of host machines by using recurrent neural networks. The process involved utilizing the recurrent neural networks that could accurately predict the time series data and also collect the information with flexibility. With respect to the traditional approaches and methods, this method was successful in accurate forecasting and gave better outcomes.

Liu et al. [22] propose A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning. It stood for a novel hierarchical framework that could address and solve all the possible power and resource allocation problems in the cloud based platforms. The proposed system took into account the virtual machines servers and various resources. The rising use of the reinforced deep learning solutions also helped in restructuring the entire concept and model. The workload prediction could be used for several other formats and henceforth is great way to rebuild the systems.

Zuo et al. [23] proposed A Multiqueue Interlacing Peak Scheduling Method Based on Tasks Classification in Cloud Computing. It was mainly a scheduling scheme that was further improved. The resource allocation and tasks classification was carried out on the basis of the type of memory and the CPU consumption. The infrastructure within the workloads may vary. Put together, they give rise to a complete cloud solution. The CPU specific tasks were classified differently and the peak scheduling was used for it. The interlacing was found to be useful for all the separate parts of the processing model. It could be used well with their other counterparts. Overall it was a very robust mechanism

that provided great accuracy in classification and added efficacy to the complete system.

Hu et al. [**24]** propose Three Models to Predict the Workload Based on Analysing Monitoring Data. The dataset for the cloud workload is a very important part of gauging the entire system design. The authors proposed three models for forecasting the cloud workload. And the help was taken from the dataset for the workload. By monitoring the data and information flow, it is easy to predict the workload extent and its quantity. This helps in building elasticity and also enhances the scalability of the system. The workload plays a crucial role and it must be flexible enough so that different programs can use it according to its changing requirements.

Xue et al. [25] put forth PRACTISE, a neural network based framework that could predict the future cloud workloads, peak loads etc. The cloud workload prediction has been a very active area of research and the authors primarily focused on forecasting the peak loads and their timings etc. As due to overflow of data and resources, the cloud serves hold the probability to crash and go off. So, forecasting helps in giving optimization solutions to the problems faced. This approach worked well with the methods and offered improved accuracy and elasticity.

Since the purpose of the proposed work is time series prediction, hence it is necessary to compute the required performance metrics. Since there is a chance of positive and negative errors to cancel out, hence it is necessary to compute the Mean Absolute Percentage Error (MAPE) given by

$$MAPE = \frac{100}{M}\sum_{t=1}^{N}\frac{E-E_t|}{E_t} \quad \textbf{(1)}$$

Here,
N is the total number of samples
E is the actual value
$\mathbf{E_t}$ is the predicated value

The mean square error is also evaluated often to stop training, which is given mathematically by:

$$MSE = \frac{1}{N}e_i^2 \quad \textbf{(2)}$$

Here,
E is the error
N is the number of samples

It is always envisaged to attain low error values and high values of accuracy for cloud workload prediction.

V. CONCLUSION

Cloud computing has emerged as a dominant paradigm for delivering scalable and on-demand computing resources over the internet. With the rapid growth of cloud-based applications, accurately estimating cloud performance metrics such as latency, throughput, response time, resource utilization, and availability has become critical. These metrics play a vital role in ensuring Quality of Service (QoS), efficient resource allocation, and Service Level Agreement (SLA) compliance. Traditional analytical and statistical methods often struggle to handle the dynamic and complex nature of cloud environments, which has led to the adoption of Machine Learning (ML) and Deep Learning (DL) techniques for performance estimation. Despite their advantages, ML and DL-based approaches face several challenges in cloud performance estimation. One major issue is the availability and quality of training data, as cloud environments generate massive and often noisy datasets. Additionally, deep learning models require significant computational resources and training time, which may not always be feasible in real-time scenarios. Model interpretability is another concern, especially for deep learning models, as it is difficult to understand how predictions are made. Furthermore, the dynamic and multi-tenant nature of cloud environments introduces variability that can affect model generalization. The review of various state of the art models allow for an optimized model.

REFERENCES

[1] J. Gao, H. Wang and H. Shen, "Task Failure Prediction in Cloud Data Centers Using Deep Learning," in IEEE Transactions on Services Computing, vol. 15, no. 3, pp. 1411-1422, 1 May-June 2022.
[2] D. Saxena and A. K. Singh, "A High Availability Management Model Based on VM Significance Ranking and Resource Estimation for Cloud Applications," in IEEE Transactions on Services Computing, vol. 16, no. 3, pp. 1604-1615, 1 May-June 2023.
[3] R Keller, L Häfner, T Sachs, G Fridgen,"Scheduling flexible demand in cloud computing spot markets: A real options approach", Business and Information Systems Engineering, Springer 2020, vol.62., pp.25–39.
[4] D. Kong, S. Liu and L. Pan, "Amazon Spot Instance Price Prediction with GRU Network," 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Dalian, China, 2021, pp. 31-36.
[5] D. Katayama, K. Kasai and T. Koita, "Migration Destination Selection Algorithm for Spot Instances using SPS," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 6690-6692

[6] D. Huang, L. Costero, A. Pahlevan, M. Zapater and D. Atienza, "CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds," in IEEE Transactions on Sustainable Computing, 2024, vol. 9, no. 4, pp. 661-676..

[7] Z. Amekraz and M. Y. Hadi, "CANFIS: A Chaos Adaptive Neural Fuzzy Inference System for Workload Prediction in the Cloud," in IEEE Access, 2022, vol. 10, pp. 49808-49828

[8] A. C. Zhou, J. Lao, Z. Ke, Y. Wang and R. Mao, "FarSpot: Optimizing Monetary Cost for HPC Applications in the Cloud Spot Market," in IEEE Transactions on Parallel and Distributed Systems, 2021, vol. 33, no. 11, pp. 2955-2967

[9] G. J. Portella, E. Nakano, G. N. Rodrigues, A. Boukerche and A. C. M. A. Melo, "A Novel Statistical and Neural Network Combined Approach for the Cloud Spot Market," in IEEE Transactions on Cloud Computing, 2023 vol. 11, no. 1, pp. 278-290.

[10] 6. Liu D, Cai Z, Lu Y (2019) Spot price prediction based dynamic resource scheduling for web applications. In: 2019 Seventh International Conference on Advanced Cloud and Big Data (CBD). IEEE, pp 78–83 7.

[11] Varshney P, Simmhan Y (2019) AutoBot: Resilient and cost-effective scheduling of a bag of tasks on spot VMs. IEEE Trans Parallel Distrib Syst 30(7):1512-1527

[12] Sharma P, Lee S, Guo T, Irwin D, Shenoy P (2017) Managing risk in a derivative IaaS cloud. IEEE Trans Parallel Distrib Syst 29(8):1750-1765

[13] Mishra AK, Yadav DK (2017) Analysis and prediction of Amazon EC2 spot instance prices. Int J Appl Eng Res 12(21):11205– 11212

[14] D. Huang, L. Costero, A. Pahlevan, M. Zapater and D. Atienza, "CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds," in IEEE Transactions on Sustainable Computing, 2024, vol. 9, no. 4, pp. 661-676

[15] H. Yuan, J. Bi, S. Li, J. Zhang and M. Zhou, "An Improved LSTM-Based Prediction Approach for Resources and Workload in Large-Scale Data Centers," in IEEE Internet of Things Journal, vol. 11, no. 12, pp. 22816-22829, 15 June15, 2024

[16] P Yazdanian, S Sharifian, E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction", Journal of Supercomputing, Springer 2022, vol. 77, pp.11052–11082.

[17] S Jeddi, S Sharifian, "A hybrid wavelet decomposer and GMDH-ELM ensemble model for Network function virtualization workload forecasting in cloud computing", Applied Soft Computing, Elsevier 2021, vol.88., Art.No. 105940.

[18] J. Gao, H. Wang and H. Shen, "Machine Learning Based Workload Prediction in Cloud Computing," 2020 29th International Conference on Computer Communications and Networks (ICCCN), 2020, pp. 1-9

[19] Z. Chen, J. Hu, G. Min, A. Y. Zomaya and T. El-Ghazawi, "Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning," in IEEE Transactions on Parallel and Distributed Systems, 2020, vol. 31, no. 4, pp. 923-934.

[20] L. Wang and E. Gelenbe, "Adaptive Dispatching of Tasks in the Cloud," in IEEE Transactions on Cloud Computing, vol. 6, no. 1, pp. 33-45, 1 Jan.-March 2018

[21] Martin Duggan, Karl Mason, Jim Duggan, Enda Howley, Enda Barrett, "Predicting Host CPU Utilization in Cloud Computing using Recurrent Neural Networks", 2017 IEEE.

[22] Ning Liu, Zhe Li, Jielong Xu, Zhiyuan Xu, Sheng Lin, Qinru Qiu, Jian Tang, Yanzhi Wang, "A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning", 2017 IEEE.

[23] Liyun Zuo, Shoubin Dong, Lei Shu, Senior Member, IEEE, Chunsheng Zhu, Student Member, IEEE, and Guangjie Han, Member, IEEE, "A Multiqueue Interlacing Peak Scheduling Method Based on Tasks' Classification in Cloud Computing", 2016 IEEE.

[24] Yazhou Hu, Bo Deng, Fuyang Peng and Dongxia Wang, "Workload Prediction for Cloud Computing Elasticity Mechanism", 2016 IEEE.

[25] Ji Xue, Feng Yan, Robert Birke, Lydia Y. Chen, Thomas Scherer, and Evgenia Smirni, "PRACTISE: Robust Prediction of Data Center Time Series", 2015 IEEE.