INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 09 ISSUE: 09 | SEPT - 2025 SJIF RATING: 8.586 ISSN: 2582-3930

A Survey on Machine Learning Models for Cloud Workload Prediction

Anjali Solanki[1], Prof. Pankaj Raghuwanshi[2]

***Abstract*— Cloud computing has become the backbone of modern IT infrastructure, enabling elastic resource provisioning and pay-as-you-go models. As enterprises migrate more workloads to the cloud, the challenge of predicting workload demand and managing resource utilization effectively has grown. Predictive models—especially those leveraging machine learning (ML) and deep learning (DL)—play a crucial role in ensuring that resources are allocated efficiently, costs are minimized, and performance meets Service Level Agreements (SLAs) Since cloud data is large and complex at the same time, hence it is necessary to use artificial intelligence based techniques for the estimation of cloud workload so as to improve upon the accuracy of conventional techniques. This paper presents a review on the contemporary techniques for cloud workload prediction. The performance evaluation parameters have also been discussed. Future research directions in terms of machine learning and deep learning algorithms for cloud workload prediction have been presented.**

Keywords—Cloud Workload Prediction, Aftificial Intelligence, Machine Learning, Artificial Neural Network (ANN), Mean Absolute Percentage error, Mean Square Error.

I. INTRODUCTION

Cloud Computing has revolutionized computational technology with cloud based platforms catering to the needs of systems unable to run complex processes on available hardware. The basic services provided by cloud computing are [1]

1) PAAS: Platform as Service
2) IAAS: Infrastructure as Service
3) SAAS: Software as service

With more sophisticated applications, it has become mandatory for tech giants to resort to cloud based services [2].

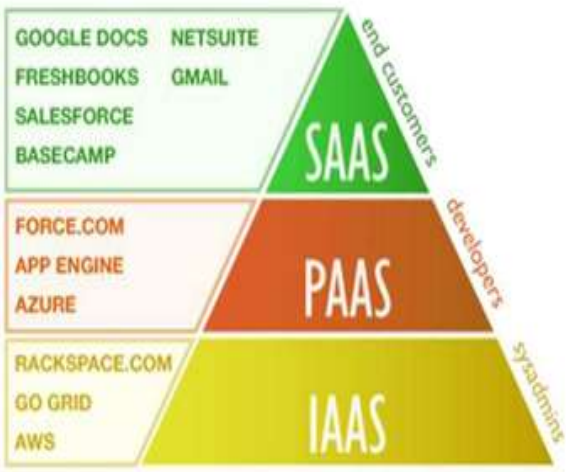

Fig.1 Cloud Services

With increasing number of users as well as large data sized, cloud workload has also seen a surge. Hence it is necessary to forecast cloud workload since several users try to access cloud services. However, the data being large and complex needs the aid of Artificial Intelligence for the prediction for the prediction purpose [3]. Cloud workload forecasting is typically challenging due to the number of users and the enormity of the data [4].

II. MACHINE LEARNING AND DEEP LEARNING FOR PREDICTING CLOUD WORKLOADS

Artificial Intelligence and Machine Learning (AI &ML) are preferred techniques for analyzing large and complex data. Generally, artificial neural networks (ANN) are used for the implementation of artificial intelligence practically [5]. The architecture of artificial intelligence can be practically implemented by designing artificial neural networks. The biological-mathematical counterpart of artificial neural networks has been shown below [6].

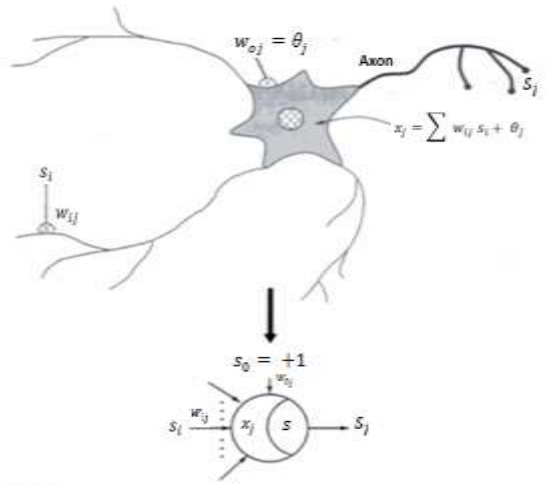

Fig.1 Biological-Mathematical Counterpart of ANN

The mathematical conversion of the ANN can be done by analyzing the biological structure of ANN. In the above example, the enunciated properties of the ANN that have been emphasized upon are [7] –[8]:

1)Strength to process information in parallel way.
2) Learning and adapting weights
3)Searching for patterned sets in complex models of data [9].

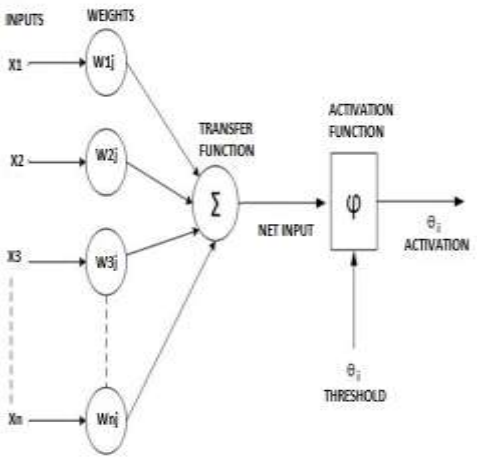

Fig.2 Mathematical Modeling of ANN

To see how the ANN really works, a mathematical model has been devised here, to indicate the functions mathematically [10]. Here it is to be noted that the inputs of information parallel goes on into the input layer as specified whereas the end result analysis is marked from the output layer [11].

The feature of parallel acceptance and processing of data by the neural network serves a vital role. This ensures efficient and quicker mode of operation by the neural network. Also adding to it, the power to learn and adapt flexibly by the neural network aids in processing of data at a faster speed [12]. These great features and attributes make the ANN self dependent without requiring much intervention from humans. The output of the neural networks can be given by [13]:

$$Y = f(\sum_{i=1}^{n} X_i . W_i + \theta_i) \quad (1)$$

Here,
Y represents output
X represents inputs
W represents weights
Θ represents Bias
f represents the activation function

Training of ANN is of major importance before it can be used to predict the outcome of the data inputs. Neural Networks can be used for a variety of different purposes such as pattern recognition in large and complex data pattern sets wherein the computation of parameters would be extremely daunting for conventional statistical techniques [14]

III. PREVIOUS WORK

This section highlights the prominent work in the domain.

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. [26] 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.

Abdelwahab et al. [26] Enabling Smart Cloud Services through Remote Sensing: An Internet of Everything Enabler survey. The concept of remote sensing has been utilized in this research work. The use of the cloud services by the IoT and remote sensing techniques has been the subject of the study. It is a very good concept to use both the technologies together. Already the emergence of the IoT has been coupled with the cloud based services and their amalgamation has been quite a success. This survey points out the aspects of the remote sensing for cloud services and other allied areas where it can be applied. Cloud based platforms provide several applications such as web services, security services, big data and machine learning services.

This section presents a comprehensive review of the different previous approaches

Authors	Approach
Yuan et al. [15]	Proposed VSBG which 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.
Yazdanian et al. [16]	A deep learning based Long Short Term Memory Based approach for cloud workload forecasting.
Jeddi et al. [17]	Group Method of Data Handling-Extreme Learning Machines (GMDH-ELM) ensemble neural network used for cloud workload prediction.
Gao et al. [18]	A machine Learning based trace-driven experiments based on Google cluster trace demonstrates that our clustering based workload prediction methods outperform other comparison methods and improve the prediction accuracy to around 90% both in CPU and memory.
Chen et al. [19]	The approach used for cloud workload prediction used in this paper is neural networks in conjugation with differential evolution approach
Wang et al. [20]	This approach focusses on the dispatch of tasks and load on the cloud server as an optimization problem
Duggan et al. [21]	The proposed approach presents a forecast of CPU utilization of Cloud servers using recurrent neural network learning.
Liu et al. [22]	This approach uses a reinforcement learning in neural networks for resource management in cloud servers.
Zuo et al. [23]	The approach uses a multi queue based interlacing approach for classification of cloud computing services.
Hu et al. [24]	The approach focusses on cloud workload estimation and the performance metrics is accuracy. The approach tests the model for different data sets.
Xue et al. [25]	The proposed approach predicts the number of data centers in cloud based servers based on the analysis of previous data.
Abdelwahab et al. [26]	The approach proposed an internet of everything based approach using cloud based services to connect devices over internet

Table.1 Comparative Analysis of Previous Work

Traditional machine learning models such as Linear Regression, Support Vector Regression (SVR), Random Forests, Gradient Boosting Machines (GBM), and XGBoost have been successfully applied for predicting cloud workloads. These models analyze historical workload metrics—such as CPU usage, memory consumption, and network bandwidth—to identify patterns and trends [27]. They are computationally efficient and interpretable, making them suitable for scenarios where quick predictions are required. However, their ability to capture highly complex temporal dependencies is limited compared to deep learning approaches [28].

Deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), Convolutional Neural Networks (CNNs), and Transformer-based architectures are capable of learning intricate temporal and spatial relationships in workload data. LSTM and GRU models are particularly effective in handling long-term dependencies, while attention-based models and Transformers improve performance by focusing on critical time points in the workload sequence. These models can adapt to evolving workload patterns, making them more robust in volatile cloud environments [29]. While ML models are simpler and faster to train, DL models often provide superior accuracy for complex and large-scale datasets [30]. For example, Random Forest and XGBoost perform well when the workload data exhibits structured but moderately complex patterns. In contrast, LSTM and Transformer-based models excel in scenarios where there are non-linear interactions and multi-step dependencies in resource demand. Hybrid approaches that combine ML's interpretability with DL's high accuracy are emerging as promising solutions [31].

IV. PERFORMANCE METRICS

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 [32]:

$$MAPE = \frac{100}{M}\sum_{t=1}^{N}\frac{E-E_t|}{E_t} \quad (2)$$

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:

$$MSE = \frac{1}{N}e_i^2 \quad (3)$$

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.

CONCLUSION

It can be concluded that cloud workloads are inherently volatile, often influenced by user demand spikes, time-of-day effects, and seasonal variations. These non-stationary patterns can lead to concept drift, where the statistical properties of the workload data change over time. Traditional ML and DL models trained on static datasets may fail to adapt to these evolving trends, resulting in inaccurate predictions and inefficient resource allocation. Predicting workload and resource utilization requires analyzing multiple variables, such as CPU usage, memory consumption, I/O operations, network traffic, and application-specific metrics. The high dimensionality of these features can cause overfitting in ML/DL models, especially when training data is limited. Effective feature selection or extraction is challenging because important predictors may not always have obvious correlations with the target workload behavior. This paper presents a comprehensive review of machine learning and deep learning models for predicting cloud workload and resource utilization.

REFERENCES

[1] S Afzal, G Kavitha, "Load balancing in cloud computing–A hierarchical taxonomical classification", Journal of Cloud Computing, Springer 2019, vol.8, no.22, pp.1-24.

[2] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su'ud and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," in IEEE Access, 2020, vol. 8, pp. 130500-130526.

[3] M. Ala'anzy and M. Othman, "Load Balancing and Server Consolidation in Cloud Computing Environments: A Meta-Study," in IEEE Access, 2019, vol. 7, pp. 141868-141887.

[4] J. Bi, H. Yuan, M. Zhou and Q. Liu, "Time-Dependent Cloud Workload Forecasting via Multi-Task Learning," in IEEE Robotics and Automation Letters, 2019, vol. 4, no. 3, pp. 2401-2406,

[5] L Bao, J Yang, Z Zhang, W Liu, J Chen, "On accurate prediction of cloud workloads with adaptive pattern mining", Journal of Supercomputing, Springer 2023, vol.79, pp. 160–187.

[6] C Yang, Q Huang, Z Li, K Liu, F Hu, "Big Data and cloud computing: innovation opportunities and challenges", International Journal of Digital Earth, Taylor and Francis 2017, vol.10., no.1, pp.13-53.

[7] A. Rossi, A. Visentin, S. Prestwich and K. N. Brown, "Bayesian Uncertainty Modelling for Cloud Workload Prediction," 2022 IEEE 15th International Conference on Cloud Computing (CLOUD), Barcelona, Spain, 2022, pp. 19-29.

[8] Jitendra Kumar , Ashutosh Kumar Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution", Future Generation Computer Systems, Elsevier 2018, vol.81, pp.41-52.

[9] D Saxena, AK Singh, "Auto-adaptive learning-based workload forecasting in dynamic cloud environment", International Journal of Computers and Applications, Taylor and Francis 2022, vol.44. no.6., pp.541-551.

[10] 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.

[11] A. K. Singh, D. Saxena, J. Kumar and V. Gupta, "A Quantum Approach Towards the Adaptive Prediction of Cloud Workloads," in IEEE Transactions on Parallel and Distributed Systems, 2022, vol. 32, no. 12, pp. 2893-2905.

[12] S Sharifian, M Barati, "An ensemble multiscale wavelet-GARCH hybrid SVR algorithm for mobile cloud computing workload prediction", International Journal of Machine Learning and Cybernetics, Springer 2019, vol.10, pp. 3285–3300.

[13] D Alberg, M Last," Short-term load forecasting in smart meters with sliding window-based ARIMA algorithms", Vietnam Journal of Computer Science, Springer 2018, vol.5, pp. 241–249.

[14] B. Feng, Z. Ding and C. Jiang, "FAST: A Forecasting Model With Adaptive Sliding Window and Time Locality Integration for Dynamic Cloud Workloads," in IEEE Transactions on Services Computing, 2023, vol. 16, no. 2, pp. 1184-1197.

[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.

[26] Sherif Abdelwahab, Member, IEEE, Bechir Hamdaoui, Senior Member, IEEE, Mohsen Guizani, Fellow, IEEE, and Ammar Rayes, "Enabling Smart Cloud Services Through Remote Sensing: An Internet of Everything Enabler", 2014 IEEE.

[27] A. Setayesh, H. Hadian, and R. Prodan, "An Efficient Online Prediction of Host Workloads Using Pruned GRU Neural Nets," arXiv preprint, Apr. 2023.

[28] A. Hadi, O. Ulah, Z. Amir, et al., "An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model," Neural Computing and Applications, 2021.

[29] S. Arbat, V. K. Jayakumar, J. Lee, W. Wang, and I. K. Kim, "Wasserstein Adversarial Transformer for Cloud Workload Prediction," arXiv preprint, Mar. 2022.

[30] Z. Zhang, D. Guo, H. Omara, et al., "Adaptive workload forecasting in cloud data centers," Journal of Grid Computing, vol. 18, no. 1, pp. 149–168, 2020.

[31] S. Karimunnisa et al., "Deep Learning-Driven Workload Prediction and Optimization for Load Balancing in Cloud Computing Environment," Cybernetics and Information Technologies, vol. 24, no. 3, pp. 21–38, Sep. 2024.

[32] J. Bi, H. Yuan, M. Zhou and Q. Liu, "Time-Dependent Cloud Workload Forecasting via Multi-Task Learning," in IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 2401-2406, July 2019.