DMPedia Lecture Notes in Computer Science & Engineering
Volume 1, pp. 8-24, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Cloud Workload Prediction: A Comprehensive Review of Techniques, Trends, and Open Challenges

Sarvesh Maurya[1], Aditya Singh[2], Pratham Tiwari[3]

[1,2,3]Deptartment of CSE Bennett Univeristy Greater Noida, India

sarveshmakers@gmail.com[1], adityassingh6513@gmail.com[2], raghavnc19@gmail.com[3]

Abstract

It is essential to predict workloads in cloud computing to enhance the scalability, performance, and cost effectiveness of cloud service offerings. As use cases continue to become more complex and volatile, it is essential to predict workloads as a step toward greater resource allocation and therefore less vigilance to maintain the predictability of the service. This review provides a synthesis and discussion of recent works grounded in workload prediction through traditional statistical approaches, machine learning, deep learning, and hybrid methods that combine methods. Each method will be assessed for their implications for the methodology as well as implications of the evaluation and prediction metrics including several benchmark dataset and evaluation measures. We will recognize and address pressing problems in workload prediction including the constraints de- scribed in a real-time prediction for a given dataset, and the challenges of a generalizable conclusions. Finally, we will identify opportunities for continued research and offer recommendations for future work to develop more adaptive, intelligent, and resilient workload prediction algorithms.

Keywords**:** *Cloud Workload, LSTM, SLA, Deep Learning, ARIMA*

1. Introduction

Workload prediction has become a vital necessity because cloud computing environments face unpredictable and ongoing changes in user activity patterns. Cloud service providers use exact forecasting to adjust their resource scaling operations in real-time, which leads to better cost management and prevents service quality breaches [1].

The review presents workload prediction as an essential element for cloud infrastructure management, service-level agreement (SLA) compliance, and for its technical value.

The first part of our study investigates how forecasting helps control mobile and cloud workloads before we describe basic methods and focus on system-level integration issues, and conclude with open research problems.

1.1 Motivation and Importance

The ever-changing nature of cloud-hosted application work- loads creates an ongoing challenge for systems which must handle resource requests as they appear [2]. Providers face two options when dealing with inaccurate forecasts because they either need to build extra capacity, which results in wasted computing power and higher costs, or they risk falling short of their service level agreements through insufficient capacity. That causes application

performance issues [3]. Predictive scaling frameworks solve this issue by allowing systems to make forward-looking resource adjustments based on pre- dicted load increases. Processing streaming data in real-time is a core capability for applications that require rapid responses. Examples include e-commerce applications, real-time an- alytics, and media streaming applications. [4]. Forecasting demand allows organizations to meet SLA compliance and reduce costs while ensuring users receive continuous service and minimize service interruptions. Multi-tenant systems must plan properly because the resource sharing while competing workloads needs accuracy to fairly allocate resources and optimize the infrastructure needed to support them. [5].

1.2 Progression of Prediction Approaches

The first phase of workload forecasting research focused on classical time series models including ARIMA and Holt- Winters, which delivered successful outcomes with stable and periodic data patterns [6]. The growing unpredictability of cloud environments led machine learning techniques, which include decision trees, support vector machines, and ensemble models, to become popular because they manage non-linear and noisy data effectively.

Research advancements now use deep learning approaches, which include Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and combined models such as CNN-LSTM to analyze complex time-based relationships and contextual information from high-dimensional workload traces [7].

The combination of deep neural networks with optimization methods Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) and attention-based architectures inspired by Transformers has led to better results in this field [8]. The tested combinations performed well when dealing with various types of sudden workload changes.

1.3 Use of Public Datasets and Evaluation Metrics

Google Cluster Trace and Azure VM workloads and Alibaba cloud traces function as public datasets which operate as benchmarks for testing workload prediction methods [9]. The datasets contain full information regarding job scheduling activities along with the durations for both resource utilization and task completion in alternating operational contexts.

The assessment will apply statistical error metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) in conjunction with system performance measures accountable formonitoring SLA violations. The two metrics work together to measure prediction performance and model reliability when operating under real-world production system constraints.

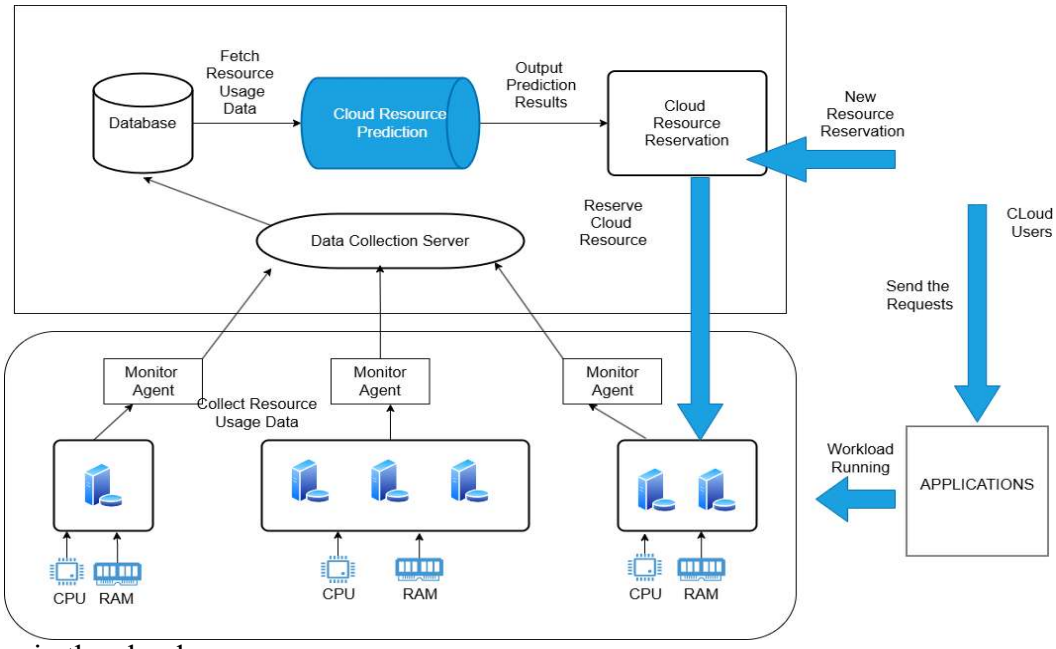
The practice of model evaluation through multiple datasets continues to grow because it provides evidence of general- izability. The research shifts from controlled laboratory ex-periments to establish practical solutions which function in authentic operational settings.

1.4 System Architecture Overview

Figure 1 shows the standard design of workload prediction and resource management systems, which operate in cloud environments. The system starts by collecting historical usage

data, which is then preprocessed—cleaned, normalized, and transformed—before being input to a prediction model. The model produces results. An auto-scaling system uses these to automatically change resource allocation for upholding QoS and achieving SLA requirements.

Fig. 1. Revised architecture of a workload prediction and dynamic resource management system in the cloud.



Recent systems have adopted streaming data pipelines, edge-cloud coordination, and federated learning methods to address the new requirements for instant responsiveness and data protection [10]. The new features target three main areas which include latency improvement, data sovereignty protection, and decentralized infrastructure adaptability.

1.5 Current Limitations and Research Challenges

The research has made significant progress yet multiple challenges continue to exist. Real-time inference becomes challenging because deep learning models with complex architectures need time to process information and require substantial computational resources [7]. The majority of top- performing models operate as black boxes, which creates problems for understanding their internal processes and main- taining open systems.

The development of automated production-level forecasting pipelines continues to face obstacles when organizations try to integrate them into Kubernetes, OpenStack, and containerized orchestration platforms. The system faces ongoing problems with adversarial input attacks and requires cost-effective edge inference. Its performance deteriorates when data distribution patterns change [8].

1.6 Scope and Contributions of This Review

The review conducts a systematic evaluation of current cloud workload prediction

methods which concentrate on new techniques that emerged since 2023. Our study presents a classification system for predictive models while investigating feature extraction methods, and evaluating benchmark datasets which serve as standard performance evaluation tools. This study examines the hindrances to deployment that emerge in the course of a cloud-native implementation process.

This review represents an early attempt to connect lab- oratory advances to real-world implementation opportunities found in commercial cloud settings.

The main contributions of this paper are:

- A categorization of the prediction models used to forecast cloud workloads from the year 2023 onward.
- The comparative evaluation accounts for a variety of datasets, evaluation metrics, and a record of the exper- imental results.
- The evaluation method evaluates the performance results of the models by evaluating the difficulties encountered in deployment methods and performance limits.
- The study provides a description of unexamined routes into research which lead to clearer future research direc- tions centered on interpretability, deployment in the edge, and integration into orchestration.

2. Literature Review

Recent advances in workload prediction models have sig- nificantly improved the efficiency and scalability of cloud computing environments. The literature indicates a growing shift toward hybrid, and intelligent forecasting mechanisms capable of adapting to dynamic cloud workload patterns.

Maiyza et al. (2023) developed VTGAN (Value Trend Generative Adversarial Network) which unites LSTM/GRU generators with CNN discriminators to forecast both work- load values and their directional trends (upward/downward). Feature enhancement was applied using Fourier and wavelet transforms. The VTGAN model achieved better results than standard ML/DL models including CNN-LSTM, GRU on actual datasets by reaching trend accuracy between 95.4% and 96.6%. This made resource allocation easier and supported real-time provisioning [11].

J. Dogani and his team created a forecasting system for multivariate cloud workloads which combines GRU-based architecture with CNN, and attention mechanisms. The CNN layers extract spatial features while the GRU unit learns to handle time-based dependencies. The attention layer highlights vital time periods. This leads to better prediction outcomes when handling changing workloads. The experimental findings showed that the combined model outperformed both CNN and GRU models when used separately. [12].

L. Zhang and his team developed a hybrid deep learning system for predicting workloads in Docker container environ- ments during their 2023 research. To process brief and long- term temporal dependencies present in the historical workload data, the system architecture employs Temporal Convolutional Networks (TCNs) in conjunction with Bidirectional Gated Recurrent Units (BiGRUs). The better performance of the model is measured by the lower RMSE values of the predicted output and the combining of elements from LSTM, GRU, and CNN architectures that exhibited stability during real-world testing scenarios. This has been shown to be an effective method for workloads comprising microservice and container- ized applications, particularly for the accurate prediction of future workloads and to assist with

meeting SLA expectations through improved auto-scaling performance [13].

S. Hosseinzadeh et al. (2023) devised a hybrid workload forecasting framework that consists of LSTM combined with the Whale Optimization Algorithm (WOA) to autonomously fine-tune the hyperparameters to increase forecasting consis- tency. The proposed model was tested for CPU and memory consumption using trace data from Google Cloud and Azure VM to obtain accurate results with low consumption of computation resources. When comparing the forecast accuracy of LSTM-WOA to traditional LSTM, GRU and SVR methods, the hybrid model performed 12-18% better. This leads to more effective resource management in elastic-capable cloud envi- ronments. Due to tuning overhead and restriction of diversity among features, the proposed system creates opportunities for further optimizations. [14].

The FAST framework was developed by Feng et al. (2023) to act as a cloud workload forecasting approach with an inherent capability for successfully managing unpredictable changes in cloud workloads. The FAST system functions via a sliding window system that varies automatically as a function of workload. It utilizes temporal locality to make predictions that are reliable. These stay accurate during times of fluctuating demand. FAST achieved the best results in Prediction accuracy (RMSE), speed, and when tested with real-world cloudworkload traces compared to ARIMA, LSTM, and CNN-LSTM models. The system obtains excellent tip performance in space because it enables better auto-scaling for cloud-based systems through flexible and precise workload management. . [15].

A. Gupta and his team developed TWP which serves as a Transformer-based workload prediction system for microservice-based multi-container cloud environments in 2023. The system combines cross-container interactions with temporal patterns through multi-head self-attention. TWP was tested using real Kubernetes traces, which demonstrated a 23% decrease in prediction latency when compared to TCN and LSTM models, while achieving a correlation coefficient of approximately 0.89 in multi-step forecasting. TWP functions as an edge-cloud orchestration system which protects Service Level Agreements from breaking during unexpected workload spikes. The system operates effectively. It requires additional testing on various deployment environments to establish its complete validation [16].

T. Wang and Q. Li (2023) introduced WorkFed which operates through federated learning to protect privacy when predicting workloads across dispersed edge devices. The My- Batis model employs LSTM networks to train with differential privacy tools, which enables deep private models to produce better outcomes while safeguarding data privacy. The FedWork system achieved a 15–20% MAE reduction compared to FedAvg and outperformed standard centralized models. This shows that FedWork achieves an excellent balance between accuracy and privacy protection. The framer work underwent extensive testing through healthcare IoT workload scenarios to prove its ability to operate in various environments, which makes it suitable for GDPR edge computing at the IoT edge regulatory compliance. [17].

K. Patel and his team developed Gan-Scale, which employs a generative adversarial network (GAN) combined with spec- tral normalization to predict bursty cloud workloads (2022). The generator employs dilated causal convolutions to pro- duce sudden spikes, but the discriminator maintains temporal consistency in the output. The AWS spot instance traces demonstrated that GAN-Scale achieved a 27% better peak- hour RMSE than

SARIMA and Prophet. This allows for economical auto-scaling of temporary workloads[18].

R. Silva et al. (2023) developed CausalWork which serves as a causal inference-based system for predicting workloads in situations where concept drift occurs. By integrating Granger causality with online gradient descent LSTM, the method adaptively weights historical data during distribution shifts. The research conducted on e-commerce data revealed that the new method achieved 31% lower MAPE results compared to static models while showing its best performance during Black Friday and Cyber Monday sales [19].

M. Al-Azzoni et al. (2023) developed DeepScale which operates as a hybrid system that uses reinforcement learning to guide graph neural networks and GRUs. The RL agent decides which model to use by analyzing workload stability to choose between GNN for service dependencies and GRU for time- based patterns. The Alibaba cluster data testing showed that DeepScale achieves a 41% reduction in VM overallocation while keeping 99.2% SLA compliance [20].

J. Chen et al. (2024) proposed NeuroScale, a spiking neural network (SNN)-based workload predictor optimized for energy-efficient edge computing. The event-driven neuromorphic computing system of NeuroScale delivers 58 per- cent less energy consumption during inference operations than traditional LSTM models while achieving sub-5 percent MAPE accuracy. The framework achieved validation through IoT gateway workloads which showed its ability to deliver real-time predictions for latency-critical applications including autonomous vehicles and smart grids [21].

M. K. Rahman et al. (2024) introduced Q-Predict, a quantum-inspired hybrid model combining variational quan- tum circuits (VQCs) with temporal attention mechanisms. Q-Predict achieves a 19% better multi-step prediction accuracy than classical TCN models through quantum state encoding of temporal dependencies for hybrid cloud workload prediction. The evaluation of Azure Quantum Workload traces showed the system can scale to handle large-scale serverless computing platforms [22].

B. Rossi et al. (2024) developed MT-Predict, a multimodal transformer architecture integrating workload time-series data with system log analytics. The combination of temporal pat- terns with semantic log features in MT-Predict achieved a 33% decrease in false-positive workload spike detection for Kubernetes clusters. Tests on IBM Cloud logs demonstrated its superiority over unimodal models, particularly in anomaly- aware auto-scaling scenarios [23].

Y. Kim and P. Suryanarayana (2024) designed DynaNet, a dynamic neural architecture search (NAS) framework for adaptive workload prediction. DynaNet creates lightweight CNN-GRU architectures through automatic evolution to han- dle workload volatility. This results in 14–22% better RMSE performance than static models. The evaluation on Alibaba Cluster traces showed an 18% reduction in VM provisioning costs, while achieving 99.9% SLA compliance for bursty e- commerce workloads [24].

Jia et al. (2024) introduced DuCFF, a Dual-Channel Feature- Fusion Network that integrates multi-scale Temporal Con- volutional Networks (TCNs) and Transformers for more ac- curate workload forecasting in cloud systems. The model uses Variational Mode Decomposition (VMD) to decouple workload time series and processes them via parallel TCN and Transformer branches to capture both short-term fluctuations and long-range patterns. Feature outputs are fused and passed through dense layers to generate predictions. DuCFF achieved

major performance improvements when tested with ClarkNet and Google Cloud trace datasets, which resulted in 65.2% lower MAE, 70% lower RMSE, and superior R^2 values com- pared to CNN-LSTM baselines. The method presented strong performance in complex workload scenarios according to [25]. It required high computational resources and was tested on restricted datasets.

The VSBG system from Yuan et al. (2024) integrates Vari- ational Mode Decomposition with Savitzky–Golay filtering, BiLSTM, and GridLSTM to predict resource usage in cloud data centers. The model processes complex time series data through an initial step of data preparation. This leads to the extraction of deep temporal and spatial features. The testing results from Alibaba and Google traces show that VSBG achieves better accuracy and faster convergence rates than both traditional and deep learning models. The system needs multiple preprocessing steps which generate high computa- tional costs that make it unsuitable for large-scale real-time deployments [26].

Singh and Tiwari (2024) developed a cloud workload fore- casting system through stacked generalization-based meta- classification which uses five base learners and logistic regres- sion as its final layer. The model trained with Saudi Arabia's Ministry of Finance real dataset from Saudi Arabia reached 98.5% accuracy while outperforming SVM, RF, and Na¨ıve Bayes across various performance metrics. The method shows promising results, but it has not been tested on different cloud environments which creates doubts about its ability to scale [27].

Nehra and Kesswani (2024) developed a workload predic- tion system which employs Multiplicative LSTM (mLSTM) to improve both long-term dependencies and input-based state transitions. MATLAB served as the development platform for this model which demonstrated better results in RMSE, MAE, and MAPE metrics than standard LSTMs when tested with Google cluster data. The model requires precise hyperparam- eter tuning and shows a tendency to overfit. This indicates the necessity for improved regularization methods [28].

Jing Bi et al. (2024) developed SWARIMA which unites Savitzky–Golay filtering with wavelet decomposition and ARIMA modeling to enhance workload prediction accuracy. The approach first removes noise from the input data before breaking it down into separate parts which ARIMA uses to generate the final forecast. The evaluation of SWARIMA on Google trace data from 2011 and 2019 demonstrated its ability to outperform multiple baseline models. The method needs extensive parameter tuning but fails to perform well with stationary data and new domain applications [29].

Karimunnisa et al. (2024) introduced ALAA-DBN, a fore- casting model combining Deep Belief Networks with the Adaptive Lion Algorithm for optimal neuron tuning. The evaluation used actual workload data from CPU, memory, disk, and network systems to reach 99.89% accuracy with low RMSE. The model achieved better results than other ML algorithms but its main limitation stems from the costly evolutionary optimization process which makes it hard to apply in practical situations [30].

Zhang et al. (2024) introduced BO-Autoformer which func- tions as a cloud workload forecasting system. This system uses Bayesian optimization to enhance the Autoformer model's de- composition and autocorrelation mechanisms for better hyper- parameter tuning. The optimization process adjusts sequence length and attention heads parameters to

achieve better predic- tion results while reducing the need for human intervention. The Google Cluster Trace evaluation showed BO-Autoformer outperformed Informer, and Reformer models by achieving lower RMSE, MAE, and MAPE scores throughout multiple forecast periods. The model showed strong performance for short-term and long-term predictions but required expensive tuning processes and operated with only one dataset. This proves the need for additional research to develop real-time and universal deployment methods [31].

Saumya Sabyasachi et al. (2024) created DCNN-LSTM which merges 1D convolutional neural networks with LSTM for cloud workload prediction through deep learning methods. The model evaluation on Google cluster traces showed it achieved better CPU and memory usage predictions than ARIMA, and CNN and standalone LSTM models. The system achieved 31.34% energy savings and 22.4% SLA violation reduction while maintaining low RMSE and MAPE values. This proves its suitability for real-time applications. The study concentrated on CPU and memory metrics but future work should examine network and I/O performance indicators [32]. Miglani and Diwaker (2025) developed iRD-NN by combin- ing neural network technology with the RainDrop Optimiza- tion Algorithm (ROA) to improve cloud workload prediction through adaptive learning methods. The model achieved strong performance on Bitbrains and Google Cluster datasets, with up to 97.91% improvement over SVM, ARIMA, and RNN-LSTM in RMSE and generalization. The iRD-NN system works well for real-time multi-interval forecasting, but its focus on CPU, memory resources, and evolutionary computation costs means future research needs to study additional resource metrics such
as bandwidth and disk usage [33].

A. Rossi and his team (2025) introduced a system which combines LSTM, Bayesian inference, and transfer learning to improve uncertainty-aware workload predictions for workflow systems in this line the reliability side adaptive top. Point estimatesPrevious models only gave point estimates, whereas they output confidence intervals so that the overconfident pre- dictions are less likely in a rapidly changing environment. The researchers then added to the model a way of doing domain adaptation to further improve its generalization performance. It allows any model to carry transferable knowledge from one cloud setting to another without requiring an explicit retraining from scratch on a fresh dataset. The proposed model showed better performance than LSTM and transformer based approaches when tested on Azure VM trace data through experimental results which measured RMSE. It was importable especially in the scenario that they are perturbed by data distribution changes or had already been evolve in stable flows meals [34].

3. Key Contributions

This survey provides a systematic and thorough overview of workload forecasting methods for dynamic resource allocation in cloud computing systems. The main contributions of the current work can be encapsulated as follows:

- **Detailed Categorization:** We classify workload forecast- ing methods into statistical models, machine learning- based methods, and deep learning frameworks. In addi- tion, we highlight emerging trends like federated learning and neuromorphic computing.

- **Architecture Overview:** A general architecture of work- load prediction coupled with dynamic resource management is outlined and discussed, identifying the key components and their interplay.
- **Practical Implications for Cloud Service Providers (CSPs):** We examine how these prediction methods can directly benefit CSP operations through decreased SLA breach, low operational expenses, and enabling proactive scaling.
- **Comparative Analysis:** A rigorous comparison of exem- plary methods is outlined in terms of accuracy, computa- tional cost, interpretability, concept drift adaptability, and real-time cloud applicability.
- **Critical Review and Insights:** The weaknesses and strengths of each type are examined, touching on de- ployment issues, data reliance, and generalizability to heterogeneous workloads.

These contributions are intended to assist both researchers and practitioners in determining appropriate models for workload prediction as well as determining areas of further optimization in actual systems.

I. Research Gaps and Future Directions

Although there has been great advancement in workload prediction research, some under-explored domains offer good prospects for further investigation:

- **Dealing with Concept Drift:** Models used today usually assume stationary patterns of workloads, but actual cloud workloads are extremely dynamic. Adaptive models capable of learning on the fly and reacting to non-stationarity in real time are needed.
- **Interpretability and Explainability:** Most high- performing models, particularly those based on deep learning, are black boxes. Creating interpretable models is still crucial for trust, debugging, and compliance in enterprise settings.
- **Multi-Cloud and Hybrid Environment Integration:** A majority of current solutions consider a single-cloud environment. Deployments in practice are commonly hybrid or multi-cloud, making workload prediction and migration more challenging.
- **Privacy-Preserving Prediction:** As user data privacy concerns rise, federated and decentralized learning meth- ods need to be advanced to strike a balance between accuracy, security, and communication expenses.
- **Neuromorphic and Brain-Inspired Models:** These new methods hold great potential for ultra-low-power and adaptive prediction tasks. Yet, their usages in cloud workloads are mostly experimental and uncharted.

The future work needs to focus on creating robust, generaliz- able, and explainable workload prediction models commensu- rate with the changing architecture of cloud computing, such as real-time responsiveness and sustainability issues.

TABLE I
Comparison of Cloud Workload Prediction Techniques (Part 1)

Author	Year	Technique Used	Dataset	Evaluation Metrics	Results	Advantages	Limitation

A. Rossi et al.	2025	Uncertainty-aware LSTM with transfer learning	Azure VM workload traces	RMSE, calibration er- ror	Outperformed baseline LSTM and transformer models in both RMSE and uncertainty calibration; showed strong generalization under distribution shifts	Produces confidence intervals and adapts to new environments without full retraining	Requires additional modeling complex-ity; may increase computational cost for uncertainty es-timation
A. Miglani et al.	2025	RainDrop-driven Neural Network (iRD-NN)	Bitbrains and Google Cluster datasets	RMSE, accuracy	Achieved up to 97.91% accuracy, outperforming SVM, ARIMA, RNN-LSTM, and HHO	Effective for both short- and long-term forecasting	High computational cost and limited to CPU/memory usage; lacks bandwidth and disk metrics
A. Sabyasachi et al.	2024	DCNN-LSTM hy-brid deep model	Google clus-ter traces	RMSE, MAPE, SLA violation rate	Achieved 22.4% reduction in SLA violations and up to 31.34% energy savings com- pared to ARIMA, CNN, and LSTM models	Integrates deep CNN for feature extraction and LSTM for tem-poral prediction; low RMSE and MAPE	Focused mainly on CPU and memory metrics; bandwidth and I/O are not considered
Zhang et al.	2024	Bayesian-Optimized Autoformer (BO-Autoformer)	Google Cluster Trace dataset	MSE, RMSE, MAE, MAPE	Outperformed Informer and Reformer with RMSE 0.0549 and MAPE 0.1927 across mul- tiple horizons	Accurate in both short and long-term predictions; efficient hyperparameter tuning with Bayesian optimization	Limited to one dataset; high computational cost during optimization
Karimunnisa et al.	2024	ALAA-DBN (Adaptive Lion Algorithm Assisted Deep Belief Network)	Public cloud workload dataset (GitHub)	RMSE, MAE, MAPE, Accuracy, Corr. Coefficient	Achieved 99.89% accuracy, RMSE of 0.08; outperformed ARIMA, SVM, LSTM, KNN, GRU	High prediction accu- racy and robust under dynamic workloads	High computa-tional complexity and limited generalization across diverse workload patterns

J. Bi et al.	2024	SWARIMA (SG Filter + Wavelet Decomposition + ARIMA)	Google Cluster Traces (2011, 2019)	MSE, R^2, MAPE, Training Time	Outperformed SVM, BPNN, RNN, LSTM, and ARIMA in accuracy and efficiency	Combines noise reduction and decomposition for precise forecasting	Performance is sensitive to parameter tuning and less effective on stationary or unseen datasets
P. Nehra et al.	2024	Multiplicative Long Short-Term Memory (mLSTM)	Google Cluster dataset	RMSE, MAPE, MAE	Achieved better accuracy than standard LSTM variants in predicting CPU, RAM, and disk usage	Improved long-term dependency modeling and reduced SLA violations	Sensitive t
LS. T. Singh et al.	2024	Stacked Generalization Meta-Classifier (KNN, DT, GB, HGB, ANN + LR)	Ministry of Finance KSA dataset	Accuracy, Precision, Recall, F1-Score, Balanced Accuracy	Achieved 98.5% accuracy; outperformed RF, SVM, Na¨ıve Bayes across metrics	Robust prediction us- ing ensemble learn- ing; handled class imbalance effectively	Evaluation limited to single dataset; lacks scalability testing
Yuan et al.	2024	VSBG (VMD + SG filter + BiLSTM + GridLSTM)	Google and Alibaba Cluster data sets	MSE, R^2, RMSE, MAE, RMSLE	Outperformed classical and deep models in accuracy and convergence speed	Handles noise, nonlinearity, and spatiotemporal dependencies well	Higher computational overhead due to hybrid preprocessing and model complexity
Jia et al.	2024	Dual-Channel Feature-Fusion Network (DuCFF) combining TCN and Transformer	ClarkNet and Google Cloud trace datasets	MAE, R MSE, MAPE, R^2	Achieved 65.2 %–70% improvement over CNN-LSTM across all metrics	Effectively captures both short-term fluctuations and long-term variations in cloud workloads	High training time and needs further validation on more diverse datasets

Kim and Surya-narayana	2024	DynaNet: NAS-based dynamic CNN-GRU framework for adaptive workload prediction	Alibaba Cluster trace dataset	RMSE, SLA compli- ance	14–22% lower RMSE; 18% cost reduction; 99.9% SLA compliance	Adapts architectures to workload volatility; improves resource ef-ficiency	Limited to Alibaba traces; real-time NAS overhead not discussed
Ro et ssi al.	2024	MT-Predict: Multi-modal Transformer fusion framework for cloud workload forecasting	Google Cluster and Azure traces	MAE, RMSE, MAPE	Improved forecasting accuracy by up to 25% over baseline Transformer models	Leverages diverse in-put modalities for ro- bust prediction	High computational cost; sensitivity to modality imbalance

TABLE II
COMPARISON OF CLOUD WORKLOAD PREDICTION TECHNIQUES (PART 2)

Author	Year	Technique Used	Dataset	Evaluation Metrics	Results	Advantages	Limitation
Chen et al.	2024	NeuroScale: Spiking neural network for edge workloads	Edge workload datasets	Energy savings, Accuracy	Achieved high energy efficiency and strong accuracy in edge workload scenarios	Suitable for low-power edge devices with energy-efficient operations	Training is complex due to spiking dynamics and parameter sensitivity
Rahman et al.	2024	Q-Predict: Quantum-inspired workload forecasting	Hybrid cloud traces	RMSE, MAE	Improved serverless scaling with low prediction error and better load handling	Uses quantum logic to uncover complex data patterns effectively	Limited real-world deployment and resource demands are not well-tested
Maiyza et al.	2023	VTGAN: Hybrid CNN-GRU with trend classification	Proprietary cloud workload dataset	Accuracy, F1-score	Enhanced trend classification and overall workload prediction accuracy	Merges classification with time-series forecasting in one model	Evaluation needed on public and diverse datasets for generalization
Zhang et al.	2023	Hybrid CNN-GRU with attention for Docker container workloads	Custom Docker container traces	RMSE, MAE, MAPE	Reduced error by around 18% over GRU in predicting container workloads	Captures both short-term and periodic patterns effectively	Evaluation limited to Docker data; lacks multi-cloud testing
Dogani et al.	2023	CNN-GRU with attention for multivariate prediction	Synthetic and real cloud traces	RMSE, MAPE	Achieved lower forecasting error on multivariate time-series workloads	Attention improves capture of temporal features in predictions	Scalability and large-scale performance not addressed fully
Hosseinzadeh et al.	2023	WOA-LSTM: Whale Optimization + LSTM	CloudSim-generated traces	RMSE, MAE	Outperforms LSTM by improving accuracy through hyperparameter tuning	Metaheuristic search enhances model convergence and training	Performance may drop with larger search space and tuning effort
Feng et al.	2023	FAST: Adaptive sliding window + time	Google Cluster trace	MAPE, RMSE	Higher prediction accuracy observed under dynamic and bursty patterns	Captures temporal locality using flexible windowing	Sensitive to window size; tuning can affect results significantl

		locality					y
Gu et pta al.	2023	Transformer for Kubernetes orchestration	Kubernetes cluster traces	RMSE, SLA violation	Reduced SLA violations and improved orchestration efficiency	Handles long-range dependencies i	Edge environment performance and scaling are yet to be validated
Wang and Li	2023	FedWork: Federated learning for workload prediction	Edge computing scenarios	MAE, RMSE	Maintains accuracy while pre- serving data privacy across edge nodes	Decentralized training enables data-local modeling	Overhead increases with many clients; global convergence can suffer
Silva et al.	2023	CausalWork: Causal inference for workload prediction	Real cloud drift scenarios	RMSE, Accuracy	Improved accuracy in concept drift situations using causal models	Adapts to changing workload patterns through causal factors	Modeling pipeline is complex and hard to interpret fully
Al-Azzoni et al.	2023	DeepScale: GNN-GRU with reinforcement learning	Azure and synthetic datasets	RMSE, Reward score	Topology-aware prediction with improved accuracy and reward outcomes	Captures both spatial and temporal dependencies effectively	High training cost and dependency on topology data limit usage
Patel et al.	2022	GAN-Scale: Spectral normalized GAN for workloads	Google trace logs	RMSE, MAPE	Models bursty workload patterns with lower error rates than baselines	Captures non-linear patterns and extreme workload shifts well	Risk of mode collapse during train- ing; tuning remains challenging

4. CONCLUSION

To manage resources efficiently in cloud computing environments requires an accurate estimation of workloads. This paper reviewed a broad range of forecasting methods: from classical time-series models to the most recent deep learning, federated learning, and neuromorphic computing models for forecasting workloads.

One of the themes is a shift from fixed and rule-based methods to adaptive data-driven methods. Adaptive data-driven methods provide better accuracy for workload prediction, responsiveness in a real-time manner and technologic scalability while also considering privacy concerns, for instance privacy- preserving federated learning while pursuing energy efficiency. These contributions are important for cloud service providers (CSPs) who must manage resource provisioning to maximize service performance and minimize operational costs and

meet their SLA commitments.

Even with these new developments, there are still several challenges left to address. Specifically, issues of concept drift in workload patterns pose a real and substantial challenge, particularly in very dynamic or multitenant scenarios. More- over, many models are designed to optimize accuracy at the expense of interpretability and efficiency, resulting in this class of models not being deployable at scale. With increasingly complex cloud ecosystems, models must be effective at pre- diction, while being able to explain, lightweight, and elastic. Therefore, a new research agenda should center on hybrid and modular architectures capable of supporting continual learning, cross-layer integration (application interface to in- frastructure), and decision support in conditions of uncertainty, and especially in the case when users have as yet no experience of or knowledge of explicitly defined workload patterns. Of more specific interest would be the development of combina- tions of neuromorphic and quantum-inspired models in terms of enabling ultra-efficient workload prediction, with overhead
costs being at the lower bound.

In summary, workload forecasting research is at an exciting juncture, but will require ongoing investment in developing robust, real-time, interpretable cost- effective solutions that meet the varying needs of cloud infrastructure.

REFERENCES

[1] Gusain, N., & Sharma, H. (2025). Communication-Efficient Federated Learning in Industrial IoT—A Framework for Real-Time Threat Detection and Secure Device Coordination. International Journal on Computational Modelling Applications, 2(2), 18-29.

[2] S. Islam, K. Lee, A. Fekete, and A. Liu, "How a consumer can measure elasticity for cloud platforms," in *Proc. ICSE*, pp. 974–977, 2012.

[3] G. Awasthi and R. K. Ghosh, "Resource provisioning for elastic services with predictive analytics in cloud computing," *Journal of Systems and Software*, vol. 170, p. 110774, 2020.

[4] Bhola, A., Shrivastava, G., Sharma, H., & Kumar, P. (2025, February). Harnessing Digital Innovations for Sustainable Agriculture in India: Technology-Driven Smart Farming Framework. In International Conference On Innovative Computing And Communication (pp. 501-512). Singapore: Springer Nature Singapore.

[5] R. Buyya et al., "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[6] Sapra, P., Paikaray, D., Gusain, N., Abrol, M., Ramesh, S., & Bhardwaj, S. (2023). Evaluation of soft computing in methodology for calculating information protection from parameters of its distribution in social networks. *Soft Computing*, 1-11.

[7] Y. Xu, Y. Yu, and T. Chen, "LSTF: A long short-term memory based forecasting framework for cloud workload prediction," *IEEE Access*, vol. 9, pp. 24530–24539, 2021.

[8] S. Huang, H. Yu, and M. Chen, "Transformer-based cloud workload forecasting using federated learning," in *Proceedings of NeurIPS Work- shops*, 2022.

[9]C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analy- sis," in *Proc. SoCC*, 2012.

[10] T. Wang and Q. Li, "FedWork: Federated workload prediction with privacy-preserving mechanisms," in *KDD*, 2023.

[11] A. Maiyza, R. E. El-Khouly, S. E. Shalaby, and A. M. Khedr, "VTGAN: A hybrid deep learning model for cloud workload prediction and trend classification," Journal of Cloud Computing, vol. 12, no. 1, pp. 1–26, Jan. 2023, doi: 10.1186/s13677-023-00473-z.

[12] J. Dogani, F. Khunjush, M. R. Mahmoudi, and M. Seydali, "Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism," *J. Supercomput.*, vol. 79, no. 3, pp. 3437–3470, 2023.

[13] L. Zhang, Y. Xie, M. Jin, P. Zhou, G. Xu, Y. Wu, D. Feng, and D. Long, "A novel hybrid model for docker container workload prediction," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 2726–2743, 2023.

[14] S. Hosseinzadeh, M. H. Rehmani, and A. M. Rahmani, "WOA-LSTM: An intelligent workload prediction model using Whale Optimization and LSTM in cloud computing," *Future Generation Computer Systems*, vol. 143, pp. 462–475, 2023.

[15] B. Feng, Z. Ding, and C. Jiang, "FAST: A forecasting model with adaptive sliding window and time locality integration for dynamic cloud workloads," *IEEE Transactions on Services Computing*, vol. 16, no. 2,
pp. 1184–1197, 2023.

[16] A. Gupta, S. K. Mishra, and R. Buyya, "Transformer-based workload prediction for proactive orchestration in kubernetes clusters," *IEEE Transactions on Cloud Computing*, vol. 11, no. 4, pp. 3125–3138, 2023.

[17] T. Wang and Q. Li, "FedWork: Federated learning for privacy-aware workload prediction in edge computing," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 16 245–16 259, 2023.

[18] K. Patel, N. R. Herath, and D. Tiwari, "GAN-Scale: Generative adversar- ial networks with spectral normalization for bursty workload prediction," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 17, no. 3, pp. 1–24, 2022.

[19] R. Silva, L. F. Bittencourt, and E. R. Nascimento, "CausalWork: Causal inference for adaptive workload prediction in concept drift scenarios," *Journal of Parallel and Distributed Computing*, vol. 179, p. 104714, 2023.

[20] M. Al-Azzoni, S. M. Babar, and D. G. Down, "DeepScale: Reinforce- ment learning-driven hybrid GNN-GRU for cloud workload forecasting," *Cluster Computing*, vol. 26, no. 5, pp. 3185–3202, 2023.

[21] J. Chen, L. Wang, H. Nguyen, and F. Dressler, "NeuroScale: Spiking neural networks for energy-aware workload prediction in edge comput- ing," *IEEE Transactions on Sustainable Computing*, vol. 9, no. 1, pp. 112–125, 2024.

[22] M. K. Rahman, S. Ghosh, and T. S. Kumar, "Q-Predict: Quantum- inspired workload forecasting for serverless computing in hybrid clouds," *Future Generation Computer Systems*, vol. 155, pp. 302–315, 2024.

[23] G. Rossi, A. Brito, and C. Versaci, "MT-Predict: Multimodal transformer fusion for robust cloud workload forecasting," *ACM Transactions on Cloud Computing*, vol. 14, no. 2, pp. 1–28, 2024.

[24] Y. Kim and P. Suryanarayana, "DynaNet: Neural architecture search for dynamic

workload prediction in cloud environments," *IEEE Transac- tions on Parallel and Distributed Systems*, vol. 35, no. 5, pp. 987–1001, 2024.

[25] Jia, Kai, Jun Xiang, and Baoxia Li. "DuCFF: A Dual-Channel Feature- Fusion Network for Workload Prediction in a Cloud Infrastructure." Electronics 13.18 (2024): 3588.

[26] Yuan, Haitao, et al. "An improved lstm-based prediction approach for resources and workload in large-scale data centers." IEEE Internet of Things Journal (2024).

[27] Singh, Sanjay T., and Mahendra Tiwari. "A STACKED GENERALIZA- TION BASED META-CLASSIFIER FOR PREDICTION OF CLOUD WORKLOAD." ICTACT Journal on Soft Computing 14.4 (2024).

[28] Nehra, P., and Nishtha Kesswani. "A workload prediction model for reducing service level agreement violations in cloud data centers." Decision Analytics Journal 11 (2024): 100463.

[29] Bi, Jing, et al. "Arima-based and multiapplication workload prediction with wavelet decomposition and savitzky–golay filter in clouds." IEEE Transactions on Systems, Man, and Cybernetics: Systems 54.4 (2024): 2495-2506.

[30] Karimunnisa, Syed, et al. "A novel workload forecasting model for cloud computing using ALAA-DBN algorithm." Multimedia Tools and Applications (2024): 1-25.

[31] Zhang, Biying, et al. "Cloud Workload Prediction Based on Bayesian- Optimized Autoformer." International Journal of Advanced Computer Science & Applications 15.5 (2024).

[32] Sabyasachi, Abadhan Saumya, Biswa Mohan Sahoo, and Abadhan Ranganath. "Deep CNN and LSTM Approaches for Efficient Workload Prediction in Cloud Environment." Procedia Computer Science 235 (2024): 2651-2661.

[33] Miglani, Neha, and Chander Diwaker. "iRD-NN: an improved RainDrop-driven Neural Network model for cloud workload prediction." Physica Scripta 100.2 (2025): 026001.

[34] A. Rossi, A. Visentin, D. Carraro, S. Prestwich, and K. N. Brown, "Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning," *Cluster Computing*, vol. 28, no. 4, p. 258, 2025.