Grenze International Journal of Engineering and Technology, June Issue

Workload Forecasting Techniques in Cloud Computing: A Review and Future Directions

Bohar Singh[1], Balwinder Kaur[2] and Anu Gupta[3]
[1]Research Scholar, Panjab University, Chandigarh, India
Email: bohar2@pu.ac.in
[2-3]Panjab University, Chandigarh, India
Email: balwinder@pu.ac.in, anugupta@pu.ac.in

***Abstract*— Workload forecasting in Cloud Computing is the process of predicting future resource demands (CPU, Memory, Bandwidth and Disk I/O). It uses historical data to optimize resource utilization and management. Accurate prediction of workload leads to lower operational costs, preventing Service Level Agreement (SLA) violations and avoiding over- or under-provisioning. Researchers have developed multiple approaches to forecast workload, broadly categorized into Statistical methods, Machine Learning (ML) and Deep Learning (DL) techniques. This review primarily discusses the significant advancements in workload forecasting within Cloud Computing environments over the past few years. It mainly focuses on ML, DL and hybrid/ensemble methods used from 2020 to 2025. The present study identifies various issues through the analysis of existing research, like dealing with non-linear workloads, high-dimensional temporal dependencies and managing data variability. The review reveals that hybrid models have superior performance compared to standalone methods. It further proposes future research directions, including transfer learning applications, multi-resource prediction frameworks, energy-aware computing and uncertainty-aware forecasting models.**

***Keywords*— Cloud Computing, Workload Forecast, Machine Learning, Deep Learning, Hybrid Learning.**

I. INTRODUCTION

Cloud Computing [1] is a system that allows individuals or organizations to access computational resources. These resources include CPUs/GPUs, storage, memory, servers, operating systems, networking and related infrastructure [2]. Many companies have been migrated to cloud data center to access these resources, which in turn has been driving high cloud demands. Cloud data centers virtualize these resources and provide services to companies in terms of Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). However, increasing virtualization and cloud migration have introduced significant challenges in resource management.

Resource management [3] involves the process of provisioning, allocation, scheduling, monitoring and deallocation of resources. It balances Quality of Service (QoS), Service Level Agreement (SLA) [4] compliance, cost and energy through methods such as optimization and Machine Learning (ML). Generally, resource management is classified into two approaches, reactive and proactive [5]. The reactive approach responds to resource changes by applying predefined rules, such as CPU or memory thresholds. It often causes resource

Grenze ID: 01.GIJET.12.2.175_20

wastage during low utilization and performance degradation during high demands. As a result, overutilization and underutilization persist as key challenges in cloud data centers. The proactive approach [6] anticipates resource needs through workload forecasting to enable preemptive adjustments. The proactive strategies address these challenges and provide accurate forecasting, smart utilization, SLA compliance and cost savings. Therefore, it is essential to predict resources in advance based on historical pattern to optimize the resource planning, efficient utilization and cost reduction.

For effective resource management, the various statistical, ML, DL and hybrid learning techniques have been examined by researchers. These techniques utilize public datasets (e.g. Google Cluster Traces, Alibaba Traces) for demand forecasting, supporting both univariate and multivariate resource predictions. However, the complex non-linear patterns, long-term temporal dependencies, multi-dimensional nature of workloads, sensitivity to outliers and hyperparameter tuning challenges make accurate prediction difficult. The progression from traditional statistical methods to ML and subsequently to DL, has significantly improved workload forecasting in Cloud Computing. Based on the analysis of papers, Figure 1 presents generalized steps used in various studies for workload forecasting in Cloud Computing. The process starts with historical workload data like Google Cluster Traces moving into the Data Preprocessing stage, where outlier detection, categorical encoding and data balancing are handled. This cleaned data goes into a Model stage where all models can be employed such as Statistical Models, Machine Learning, Deep Learning and Ensemble Learning among others. The Model Training stage consists of Hyperparameter Tuning and model validation. The final stage is the forecast of CPU, Memory, Storage and Network Bandwidth for workloads. This review article examines recent techniques developed between 2020 and 2025, provides a systematic analysis of existing workload forecasting methods, and identifies future research directions in cloud computing.

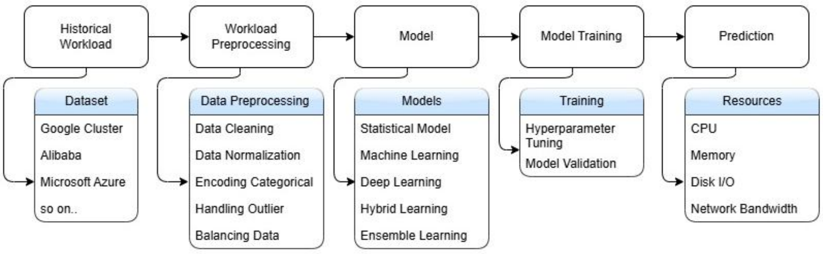

Figure 1 Generalized workload forecasting process in Cloud Computing

The paper is organized as follows: Section II discusses and analyses related work on ML, DL, Hybrid and Ensemble techniques. Section III presents a comparative summary of workload forecast techniques and datasets. Section IV highlights future research directions and Section V concludes the paper.

II. RELATED WORK

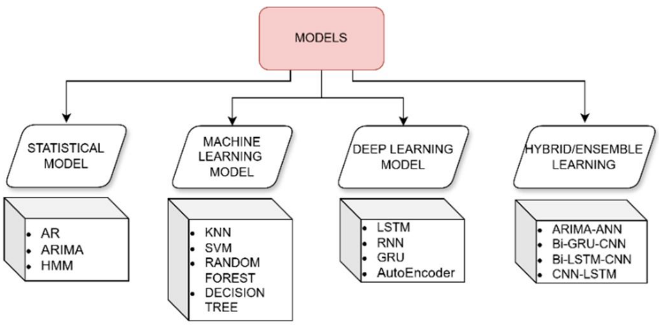

Figure 2 Workload forecasting techniques

Several researchers in their study have explored future workload prediction using ML, DL and hybrid methods. Accurate forecasting of resource demands has emerged as a paradigm for optimizing performance and resource utilization. Figure 2 represents the categories of models, from traditional to modern, used for workload forecasting in Cloud Computing. This section will provide a thorough review of the existing literature in the area of Cloud data center workload prediction. It also analyzes and compares studies on these ML, DL and hybrid methods and their unique features.

A. Machine Learning Approaches

Numerous investigations underline the application of various statistical techniques for predicting the usage of resources, such as Auto Regression (AR), which models time series based on past values, Hidden Markov Model (HMM), which captures hidden states in sequential data and Auto Regressive Integrated Moving Average (ARIMA) [7], which accounts for trends and seasonality in Cloud Computing environments. One of the big disadvantages of the aforementioned models is the fact that they failed to handle historical data with sudden changes in demand and non-linear patterns in real-world situations.

To achieve better results, previous studies have shifted their attention from traditional forecasting methods to machine learning models based on regression. In 2022, Tasheen et al. [8] examined methods such as Linear Regression (LR), Ridge Regression (RR), ARD Regression (ARDR) and ElasticNet (EN) to predict numerical outcomes, particularly metrics like CPU utilization even when workloads behave in non-linear and unpredictable ways. The research proved that the use of ML algorithms with historical data can predict future levels of the workload and the energy status of the virtual machines. The performance evaluation based on the standard metrics of Mean Absolute Error (MAE) and Mean Squared Error (MSE) showed that the ML-based solutions have made a considerable leap over the conventional statistical techniques.

Masdari and Khoshnevis [9] presented a thorough review that sorts cloud workload forecasting techniques into six major categories: regression (SVR), classifiers (random forest, SVM), neural networks (ANN, BPNN, RBF), time series (ARIMA), ensemble and emerging deep learning approaches to support proactive resource management, auto-scaling and energy optimization. The classification is significant because it deals with specific cloud issues such as dealing with bursty, non-stationary workloads and the advantages of one model over the other; for instance, SVR is suited for periodic patterns while ANN is good for non-linear workloads. The paper amalgamates results from a vast number of schemes tested on many different datasets, including Google Cluster traces, NASA HTTP, BitBrains, PlanetLab and self-collected workloads and it does so by utilizing performance metrics; MAPE, RMSE, NRMSE, execution time and cost metrics across simulators such as CloudSim and real environments like Amazon EC2.

Saxena et al. [10] presented a comprehensive approach to analyzing machine learning-based cloud workload forecasting models, sorting them out into Evolutionary Neural Networks (ENN), DL, Hybrid Learning (HL), Ensemble Learning (EL) and Quantum Learning (QL) which helps to point out the trade-off between accuracy and computational costs in different model groups. The study highlights quantum neural networks and cutting-edge ensemble/hybrid models as the leading models in terms of prediction accuracy, CPU and memory usage. The entire set of representative models was evaluated using standardized common benchmark traces (e.g. Google Cluster and PlanetLab) under variable prediction window sizes. As per the experimental outcomes, it is evident that the Evolutionary Quantum Neural Networks (EQNN) have performed the best among all the algorithms, as they presented the lowest prediction errors in all the workload traces. The performance trend thereby established was $EQNN < EL \leq HL < DL < ENN$.

ML methods can make predictions much quicker, but struggle with accuracy under high variability due to the complex nature of cloud workloads. The computational complexity of machine learning algorithms, along with training time overhead and non-linearity of modern workloads, are the main issues that make it difficult to get real-time prediction with minor SLA violation. This situation led to the exploration of DL techniques.

B. Deep Learning Approaches

Deep learning has become the dominant paradigm due to its ability to extract complex features from massive and multidimensional datasets without manual feature engineering. Deep learning architectures like Long Short-Term Memory (LSTM) networks have taken the lead in harnessing the changing nature of the workloads and temporal dependencies in workload time series data.

Bi et al. [11] introduced a DL framework BG-LSTM model that merges the bidirectional LSTM and GridLSTM architectures to detect not only the temporal relationships but also the depth-wise time–frequency features. A series of experiments conducted on long time series (2-minute slots over 29 days) compares BG-LSTM with ARIMA, SVM, standard STM, BiLSTM, GridLSTM and SG-enhanced variants.

In their technical study, Ahamed et al. [12] focused on DL based workload prediction in Cloud Computing through the investigation of RNN, MLP, LSTM and CNN models to support proactive resource management, reduction of SLA violations, over- and under-provisioning and energy waste. The results indicate that LSTM has the fewest errors all the time and the greatest capacity for handling long time-window dependencies.

Xu et al. [13] proposed a supervised deep neural network named esDNN, which applies a sliding window to convert high-dimensional multivariate cloud workload traces into time series and employs a modified Gated Recurrent Unit (GRU) to detect long-term changes and relationships in the data.

In paper [14], Saxena and Singh introduced a Multi-VM Workloads Learning-based LSTM (MVL-LSTM) model that combines the resource utilization forecasting of all the VMs on a server, aiming to enable proactive and resilient server management in the industrial clouds. The model plays a significant role as it not only replaces single-task predictors working separately by sharing a multi-task LSTM architecture, but also captures inter-VM dependencies and hence, improves failure prediction while lowering the computational overhead. The findings indicate that MVL-LSTM has continuously obtained the lowest error metrics across CPU and memory datasets and it has also provided around 15% less training time and approximately 50% less energy consumption compared to a conventional LSTM baseline.

Deep learning models such as BG-LSTM, esDNN and MVL-LSTM have strong capabilities to handle complicated temporal dependencies and non-linear patterns in cloud workloads, achieving faster training times compared to traditional LSTM baselines. However, these models suffer from high computational overhead during poor interpretability and limited generalization across diverse workload types. Hybrid approaches, which combine the computational efficiency and interpretability of statistical methods with the non-linear pattern recognition capabilities of deep learning, constitute the optimal solution for cloud workload fore-casting.

C. Hybrid and Ensemble Approaches

Several studies have explored the use of hybrid and ensemble learning for workload prediction, demonstrating superior performance over single-model approaches across diverse cloud datasets. The integration of statistical, ML and DL techniques through hybrid models provides optimized results.

In paper [6], the researcher suggested a novel two-stage ensemble framework for predicting the CPU and memory utilization. The first phase classifies the job status (Terminated, Running, Failed, Interrupted) by combining otherwise standalone models like XGBoost, CatBoost, AdaBoost and LightGBM together. The second phase, dealing with the 'Terminated' jobs only, uses combined regression models (XGBoost, Cat-Boost, AdaBoost, LightGBM) to forecast the actual consumption of CPU and memory.

Sabyasachi et al. [15] conducted a thorough investigation into the use of Deep CNN and LSTM techniques for accurate workload forecasting. The study pointed out that LSTM networks are greatly successful in dealing with sequential data and retaining long-term memory, overcoming one of the main limitations of traditional Recurrent Neural Networks (RNNs), which is the vanishing gradient problem.

Zhang et al. [16] put forward a combined CNN and BiLSTM model which is capable of predicting the load of Cloud Computing. The model utilizes CNN for extracting spatial features of workload and BiLSTM for temporal dependencies in both directions. The experiments were conducted using the Google Cluster dataset traces, wherein CNN-BiLSTM was compared against BP, LSTM, BiLSTM and CEEMDAN-ConvLSTM baselines through MSE, revealing 34-52% MSE reductions and better handling of peak loads.

Devi and Valli [17] suggested a statistical hybrid time series model that integrates ARIMA and Artificial Neural Network (ANN) and subsequently applies a Savitzky–Golay filter to estimate CPU and memory utilization in cloud data centers. The model is significant, as ARIMA captures linearity, ANN models learn nonlinearity and the filter does smoothing of forecasts into intervals, which is suitable for resource planning over longer periods than single-point estimates. Google Cluster 29-day traces and BitBrains datasets are used in the experiments to evaluate one-step and multi-step-ahead forecasts. The hybrid model was compared with ARIMA, feedforward NN, AR(30) and existing SVRT/LRT methods using RMSE, MSE, MAE, MAPE, over- and underestimation rates and it was found that the errors were consistently lower and correct prediction rates higher for both CPU and memory workloads.

Li and Zhang [18] introduced VNWSTG, a cloud workload prediction method that combines Variational Mode Decomposition, NSGA-II–optimized wavelet denoising and a hybrid Temporal Convolutional Network with self-attention plus GRU (TCN-GRU). This method was designed to manage noisy, nonlinear and non-stationary workload traces more effectively. The Google and Alibaba cluster datasets with varying sampling intervals were used for the experiments to evaluate CPU and memory utilization forecasts where-in VNWSTG was compared to GRU, TCN, TCN-GRU, Self-Attn-TCN-GRU and variants without VMD or optimized WTNR using MAE, RMSE and MAPE. The identified results show that each of the added components (VMD, wavelet denoising, NSGA-II, self-attention) contributes to the accuracy of the model incrementally and the complete VNWSTG model has the lowest prediction errors and the fastest inference among the deep baselines, consistently proving its great applicability to real-world cloud workload forecasting.

Kumar et al. [19] introduced a novel approach to forecasting VM resource requests (CPU, memory), that consists of an ensemble of several Extreme Learning Machines (ELMs) which use Blackhole algorithm-optimized voting weights for their output. This way, they can deal with the limitations of single-model approaches in coping with the diverse and quickly changing patterns of cloud workload. The proposed model is significant as it combines the fast training of ELMs with the accuracy of the ensemble and metaheuristic weighting that varies with the different traces and reduces over- and under-provisioning and energy waste in data centers. The performance of the model was evaluated by making CPU and memory demand forecasts on Google Cluster traces and PlanetLab datasets using RMSE. The throughput demonstrates high reliability across all types of workloads with a small computing cost, proving its suitability for real-time cloud resource management.

Rossi et al. [20] introduced uncertainty-aware DL model Hybrid Bayesian Neural Networks (HBNN) and probabilistic LSTMs (LSTMD) for univariate and bivariate cloud workload forecasting. These models provide predictive distributions that can be tuned to target service levels, allowing resource managers to adjust safety margins according to confidence. They are further enhanced through transfer learning to adapt across data centres of the same provider. The findings indicate that uncertainty-aware models not only enhance service-level oriented metrics and decision quality but also transfer learning on HBNN significantly increases performance when source and target datasets are from the same cloud provider, demonstrating better generalization under limited target data.

Shuvo et al. [21] proposed a hybrid LSRU model that integrates LSTM and GRU, which demonstrates better performance than both statistical models and standalone LSTM and GRU models. The LSRU model effectively predicts both short-term and long-term workloads and handles sudden workload bursts. The BHyPreC hybrid model introduced by Karim and Maswood et al. [22] combines Bi-LSTM, LSTM and GRU to achieve better performance than ARIMA, LSTM, GRU and Bi-LSTM models on BitBrains traces. The model uses sequential fusion to capture nonlinear patterns and long-term dependencies, which enables it to decrease bias that results from single-model assumptions. Maiyza et al. [2] demonstrated hybrid GAN-based models (LSTM-based and GRU-based) outperform standalone ARIMA, LSTM, GRU, Bi-LSTM and BHyPreC models. The models successfully capture long-term nonlinear dependencies in volatile traces which results in 95-96% trend accuracy and the lowest MAPE values. Devi and Valli [17] proposed a hybrid model which combines ARIMA and ANN with a Savitzky-Golay filter to demonstrate enhanced performance on Google traces and BitBrains datasets while achieving lower RMSE values. The integration reduces prediction errors and enhances model performance. The hybrid models show better performance because their modular design enables them to handle various workload patterns which include bursty, periodic and anomalous behaviors.

III. COMPARATIVE SUMMARY OF WORKLOAD FORECASTING TECHNIQUES AND DATASET

The literature provides a review of various workload forecasting techniques used from 2020 to 2025. Table I highlights the models used, their categories, key contributions, datasets utilized and performance metrics for model evaluation.

TABLE I: COMPARATIVE SUMMARY OF WORKLOAD FORECASTING MODELS

Reference	**Year**	**Model Used**	**Category**	**Dataset Used**	**Performance Metrics**	**Key Contribution**
[19]	2020	Ensemble ELMs, metaheuristic algorithm inspired by Black Hole theory	Ensemble Learning	Google cluster trace, PlanetLab	MSE, RelMAE, RMSE	Minimize operational costs and carbon footprints by accurately predicting resource demands.
[21]	2020	LSRU: LSTM and GRU	Hybrid learning	BitBrains traces	MSE, RMSE, MAE, MAPE	Achieve better accuracy over statistical models and LSTM and GRU
[11]	2021	BG-LSTM	Deep Learning	Google cluster trace	MSE, RMSE, R^2	Integrated Deep learning model for time series workload prediction
[22]	2021	BHyPreC: Bi-LSTM, LSTM and GRU	Hybrid Learning	BitBrains traces	MSE, RMSE, MAE, MAPE	Perform better than ARIMA, LSTM, GRU, Bi-LSTM
[8]	2022	LR, RR, ARDR, EN and GRU	Machine Learning and Deep Learning	BitBrains FastStorage (30 Days), Rnd (90 Days)	RMSE, provisioned (R_{CPU}, R_{memory}) and utilized (U_{CPU}, U_{memory}, $D^{th}{}_{r}$, $D^{th}{}_{w}$, $N^{th}{}_{r}$, $N^{th}{}_{t}$)	Joint workload and energy prediction for data center optimization
[13]	2022	esDNN	Deep Learning	Google cluster trace, Alibaba cluster traces	MSE, RMSE, MAE	Handle high-dimensional multivariate workload and capture long-term variance

[12]	2023	RNN, MLP, LSTM, CNN	Deep Learning	Parallel Workloads Archive	MAE, RMSE	LSTM excels in long-term time-series prediction across unbiased real-world datasets
[17]	2023	Hybrid ARIMA-ANN + Savitzky-Golay filter	Hybrid Learning	Google traces 2011 (29-day), BitBrains	RMSE, MAPE	Linear and non-linear resource prediction in cloud data centre
[2]	2023	GAN (Bi-GRU-CNN), GAN (Bi-LSTM-CNN)	Hybrid Learning	PlanetLab, BitBrains	MAPE, RMSE, MAE, Theil, ARV, POCID, R^2	Achieves 95–96% trend accuracy and lowest MAPE value
[15]	2024	Deep CNN-LSTM	Hybrid Learning	Google Cloud traces 2019	Energy Consumption, CPU Utilization, SLA Violation	Efficient multi-step workload prediction
[14]	2025	MVL-LSTM	Deep Learning	Google cluster traces (29 Days)	MAE, MSE, RMSE	Resilient multi-VM single-step forecasting for industrial server management
[6]	2025	XGBoost, LightGBM, CatBoost, AdaBoost	Ensemble Learning	Alibaba cluster traces	Precision, Recall, f1-score, R^2, RMSE, MAE, MSE	Two-stage ensemble approach with status prediction and workload prediction
[20]	2025	Hybrid Bayesian Neural Networks (HBNN), LSTM	Hybrid Learning	Google Cloud trace (2011,2019), Alibaba Cluster trace (2018,2020)	MSE, MAE, QL (Quantile Loss)	Uncertainty quantification with transfer learning for robust predictions

The analysis indicates that DL models especially the LSTM variants, outperform traditional regression methods when handling non-linear temporal dependencies. At the same time, the hybrid methods have proven to be strong and steady in different workload situations, as they make use of the advantages offered by both statistical and neural network techniques. The combination of ensemble approaches and multi-scale decomposition techniques has resulted in accuracy improvements in prediction that are incremental yet steady. All these machines learning centric models use standardize datasets for workload forecasting in Cloud Computing. Table II presents the most common datasets with their versions, sizes, focus area and limitations used for workload forecasting in cloud data center. It provides an overview of Cloud Computing datasets published by Google, Alibaba and Microsoft Azure.

TABLE II. COMMON DATASET USED FOR WORKLOAD FORECASTING

Dataset	Year/Version	Volume/Size	Focus Area	Limitations
Google Cluster Trace [23]	clusterdata-2011	Cluster of about 12.5k machines over 29 days	Task scheduling, allocation and cluster management.	Extremely large and complex dataset, high preprocessing required, no availability of network metrics.
	clusterdata-2019	Cluster of about 96k machines over 1 month		
Alibaba Dataset [24]	cluster-trace-v2017	1300 machines in a period of 12 hours	Batch workloads, online services and production DAG structures.	Small scale ~4k nodes, observation period is short, limited availability of network metrics.
	cluster-trace-v2018	4000 machines over 8 days		
Microsoft Azure Dataset [25]	AzurePublicDatasetV1-2017	VM workload containing information about ~2M VMs and 1.2B utilization readings over 30 days	Focus mainly on CPU utilization and memory usage.	No network and storage metrics available in dataset.
	AzurePublicDatasetV2-2019	VM workload containing information about ~2.6M VMs and 1.9B utilization readings over 30 days		

These datasets serve as a basis for researchers to study workload patterns, energy consumption and resource utilization at very large scales. The size of the datasets varies from thousands of physical servers to millions of Virtual Machines (VMs) and comes with a variety of features such as CPU usage, memory usage and so on. Each provider has made new versions available to indicate the increasing complexity and the larger amount of cloud data.

IV. FUTURE RESEARCH DIRECTIONS

From various studies, it has been found that the shift from conventional statistical techniques to advanced deep learning models shows the growing intricacy of cloud workload patterns. Recent studies have identified the following directions for future research:

A. Uncertainty Quantification and Transfer Learning

The current direction of research suggests that the next step would be to develop a probabilistic forecasting model that not only provides predictions but also quantifies the associated uncertainty. This would enable resource management systems to implement risk-aware allocation strategies. Additionally, there are transfer learning approaches that can borrow knowledge from well-studied workloads to improve predictions for new or less-known application types.

B. Multi-Resource and Multi-VM Prediction

The prevailing research primarily centers on metrics of single-resource, thus leading to an urgent need for multi-resource forecasting models that are full-fledged and capable of depicting the interlinkages in the utilization of CPU, memory, network and storage. Prediction systems for multi-VM workloads that account for co-located VMs can lead to better server consolidation and higher data center productivity. The incorporation of cross-resource correlation investigation into forecasting systems will pave the way for more comprehensive resource management strategies.

C. Real-Time and Adaptive Forecasting

The evolution of adaptive forecasting systems, which are capable of continuously learning from online streaming data and changing along with the workload characteristics, remains a significant challenge for researchers. The prediction in real-time with the least possible computational overhead is a key requirement for the establishment of a proactive autoscaling system. Therefore, lightweight model architectures optimized for edge computing environments could extend workload forecasting capabilities to distributed cloud environments.

D. Energy-Aware and Sustainable Computing

The future studies should establish a stronger link between the workload forecasting and energy optimization. The development of such integrated models can jointly predict resource utilization and energy consumption. The exploration of workload prediction models that are especially made for green Cloud Computing projects represents a significant path for sustainability computing research.

E. Interpretability and Explainability

Deep learning models indeed offer higher accuracy, but their black-box nature limits interpretability. The use of explainable AI approaches for workload forecasting that provide insights into prediction rationale will facilitate adoption in critical systems. The investigation into hybrid architectures that merge interpretable statistical parts with robust neural network modules could balance performance and transparency.

V. Conclusion

This review paper presents the landscape of workload forecasting techniques in Cloud Computing from 2020 to 2025. Over the years, the field has evolved significantly, moving well beyond simple statistical models and now relying heavily on advanced deep learning and hybrid/ensemble techniques. Deep learning based neural networks are more effective for capturing temporal dependencies in non-linear workload patterns, while hybrid models demonstrate robustness across diverse scenarios by combining linear and non-linear modeling capabilities. The evolution toward uncertainty-aware, multi-dimensional prediction frameworks, energy-aware computing, transfer learning and explainable AI represents the next frontier in cloud workload forecasting research.

References

[1] M. Armbrust et al., “Above the Clouds: A Berkeley View of Cloud Computing,” 2009. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[2] A. I. Maiyza, N. O. Korany, K. Banawan, H. A. Hassan, and W. M. Sheta, “VTGAN: hybrid generative adversarial networks for cloud workload prediction,” Journal of Cloud Computing, vol. 12, no. 1, Dec. 2023, doi: 10.1186/s13677-023-00473-z.

[3] M. Kumar, S. C. Sharma, A. Goel, and S. P. Singh, “A comprehensive survey for scheduling techniques in cloud computing,” Oct. 01, 2019, Academic Press. doi: 10.1016/j.jnca.2019.06.006.

[4] M. A. N. Saif, S. K. Niranjan, and H. D. E. Al-ariki, “Efficient autonomic and elastic resource management techniques in cloud environment: taxonomy and analysis,” Wireless Networks, vol. 27, no. 4, pp. 2829–2866, May 2021, doi: 10.1007/s11276-021-02614-1.

[5] F. Tahir, M. Abdullah, F. Bukhari, K. M. Almustafa, and W. Iqbal, "Online Workload Burst Detection for Efficient Predictive Autoscaling of Applications," IEEE Access, vol. 8, pp. 73730–73745, 2020, doi: 10.1109/ACCESS.2020.2988207.

[6] J. Bawa, K. Kaur Chahal, and K. Kaur, "Improving cloud resource management: an ensemble learning approach for workload prediction," Journal of Supercomputing, vol. 81, no. 10, Jul. 2025, doi: 10.1007/s11227-025-07560-9.

[7] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload prediction using ARIMA model and its impact on cloud applications' QoS," IEEE Transactions on Cloud Computing, vol. 3, no. 4, pp. 449–458, 2015, doi: 10.1109/TCC.2014.2350475.

[8] T. Khan, W. Tian, S. Ilager, and R. Buyya, "Workload forecasting and energy state estimation in cloud data centres: ML-centric approach," Future Generation Computer Systems, vol. 128, pp. 320–332, Mar. 2022, doi: 10.1016/j.future.2021.10.019.

[9] M. Masdari and A. Khoshnevis, "A survey and classification of the workload forecasting methods in cloud computing," Cluster Comput., vol. 23, no. 4, pp. 2399–2424, Dec. 2020, doi: 10.1007/s10586-019-03010-3.

[10] D. Saxena, J. Kumar, A. K. Singh, and S. Schmid, "Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313–1330, Apr. 2023, doi: 10.1109/TPDS.2023.3240567.

[11] J. Bi, S. Li, H. Yuan, and M. C. Zhou, "Integrated deep learning method for workload and resource prediction in cloud systems," Neurocomputing, vol. 424, pp. 35–48, Feb. 2021, doi: 10.1016/j.neucom.2020.11.011.

[12] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, and A. S. A. M. Al-Ghamdi, "Technical Study of Deep Learning in Cloud Computing for Accurate Workload Prediction," Electronics (Switzerland), vol. 12, no. 3, Feb. 2023, doi: 10.3390/electronics12030650.

[13] M. Xu, C. Song, H. Wu, S. S. Gill, K. Ye, and C. Xu, "esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments," ACM Trans. Internet Technol., vol. 22, no. 3, Aug. 2022, doi: 10.1145/3524114.

[14] D. Saxena and A. K. Singh, "Multi-VM-Workloads Learning-based Forecasting Model for Resilient Server Management in Industry Clouds," Procedia Comput. Sci., vol. 260, pp. 923–929, Jan. 2025, doi: 10.1016/J.PROCS.2025.03.275.

[15] A. S. Sabyasachi, B. M. Sahoo, and A. Ranganath, "Deep CNN and LSTM Approaches for Efficient Workload Prediction in Cloud Environment," in Procedia Computer Science, Elsevier B.V., 2024, pp. 2651–2661. doi: 10.1016/j.procs.2024.04.250.

[16] H. F. Zhang, J. Li, and H. R. Yang, "Cloud computing load prediction method based on CNN-BiLSTM model under low-carbon background," Sci. Rep., vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-68339-1.

[17] K. L. Devi and S. Valli, "Time series-based workload prediction using the statistical hybrid model for the cloud environment," Computing, vol. 105, no. 2, pp. 353–374, Feb. 2023, doi: 10.1007/s00607-022-01129-7.

[18] J. Li and B. Zhang, "Cloud workload prediction based on wavelet transform noise reduction and a TCN-GRU hybrid model," Cluster Comput., vol. 28, no. 12, Nov. 2025, doi: 10.1007/s10586-025-05479-7.

[19] J. Kumar, A. K. Singh, and R. Buyya, "Ensemble learning based predictive framework for virtual machine resource request prediction," Neurocomputing, vol. 397, pp. 20–30, Jul. 2020, doi: 10.1016/j.neucom.2020.02.014.

[20] 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 Comput., vol. 28, no. 4, Aug. 2025, doi: 10.1007/s10586-024-04933-2.

[21] Md. N. H. Shuvo, M. M. S. Maswood, and A. G. Alharbi, "LSRU: A Novel Deep Learning based Hybrid Method to Predict the Workload of Virtual Machines in Cloud Data Center," in IEEE Region 10 Symposium (TENSYMP), IEEE, Jun. 2020, pp. 1604–607. doi: 10.1109/TENSYMP50017.2020.9230799.

[22] M. E. Karim, M. M. S. Maswood, S. Das, and A. G. Alharbi, "BHyPreC: A Novel Bi-LSTM Based Hybrid Recurrent Neural Network Model to Predict the CPU Workload of Cloud Virtual Machine," IEEE Access, vol. 9, pp. 131476–131495, 2021, doi: 10.1109/ACCESS.2021.3113714.

[23] "GitHub - google/cluster-data: Borg cluster traces from Google." Accessed: Jan. 09, 2026. [Online]. Available: https://github.com/google/cluster-data

[24] "GitHub - alibaba/clusterdata: cluster data collected from production clusters in Alibaba for cluster management research." Accessed: Jan. 09, 2026. [Online]. Available: https://github.com/alibaba/clusterdata

[25] "GitHub - Azure/AzurePublicDataset: Microsoft Azure Traces." Accessed: Jan. 09, 2026. [Online]. Available: https://github.com/Azure/AzurePublicDataset