International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Accurate Cloud Data Center Workload Forecasting Using CEEMDAN-VMD and BiLSTM-BiGRU Networks

K. UDAY KIRAN [1], T. DREEM [2]
#1 Assistant Professor QIS College of Engineering and Technology, Ongole
#2 PG Scholar Department of MCA, QIS College of Engineering and Technology, Ongole

Abstract: Managing cloud data center resources—high-dimensional, chaotic operational data—requires forecasting workload. In order to enhance prediction without temporal repeating features in the CVCBM model, this paper presents a lightweight Bidirectional GRU (BiGRU) and Bidirectional LSTM (BiLSTM). Workload signals are concurrently denoised and decomposed in two phases (CEEMDAN and VMD). While K-Means clustering favors high-workload data, SE identifies key components. Conv1D-BiLSTM-BiGRU is a hybrid approach that acquires both short-term and long-term temporal patterns. Using input datasets, the Flask-built trained model forecasts workloads in real time. According to experimental study, the improved model offers reliable, scalable, and real-time forecasts for cloud data center resource allocation while lowering computing costs and improving forecast accuracy.

Index terms *- — workload prediction, cloud data centers, CEEMDAN, VMD, Sample Entropy, Conv1D, Bi-LSTM, BiGRU, real-time forecasting.*

1. INTRODUCTION

In today's IT architecture, cloud computing provides flexible, scalable, and on-demand computing resources (processing power, networking, and storage). It enables both individuals and organizations to effectively and dynamically access resources without having to invest a lot of money on hardware.

The usage of cloud services has grown globally, increasing the infrastructure of data centers. Google, Amazon, Alibaba, and Facebook are growing their data centers to accommodate growing compute and storage requirements [2]. Because cloud data centers use a lot of energy, warehouse-scale computers need to be high-performing and energy-efficient [3].

Because data centers are dynamic and ever-changing, management is essential. In situations with low demand and poor performance, fixed resource allocations result in underutilization. In order to maintain system consistency and meet SLAs, dynamic and intelligent work load prediction systems may proactively allocate resources [4].

One of the largest power users in data centers is the CPU. Precise workload forecasting improves cloud sustainability by lowering energy waste and increasing performance. Complex deep learning models, such as Long Short-term Memory (LSTM) networks, may learn intricate time-dependencies on

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

workload data, in contrast to linear prediction models [5].

2. LITERATURE SURVEY

a) Cloud computing load prediction method based on CNN-BiLSTM model under low-carbon background:

Industry initiatives to reduce carbon emissions have been sparked by the "double carbon" goal. Excessive carbon emissions are caused by the mismatch between load requirements and resource supply in cloud data centers, which is represented by cloud computing. This study offers a comprehensive method for forecasting carbon emissions from cloud computing. The CNN-BiLSTM model and convolutional neural network are the first steps in forecasting cloud computing demand. Real-time prediction power, which computes carbon emission forecast, is determined by the real-time prediction load of cloud computing. Develop a dynamic server carbon emission forecast model that adapts to CPU usage in order to lower carbon emissions. This study uses Google cluster data to forecast demand. The CNN-BiLSTM model predicts well, according to tests. In comparison to the multi-layer feed forward neural network model (BP), long short-term memory network model (LSTM), BiLSTM, modal decomposition, and convolution long time series neural network model (CEEMDAN-ConvLSTM), the MSE index dropped by 52%, 50%, 34%, and 45%.

b) Accurate workload prediction for edge data centers: Savitzky-Golay filter, CNN and BiLSTM with attention mechanism

In order to supply in-situ resources for workload execution, edge data centers must accurately estimate workloads. Workload is predicted using SG-CBA, a deep learning model driven by the Savitzky-Golay filter (SG filter), Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) with Attention mechanism. Workload time series are smoothed and normalized by a preprocessing program that uses an SG filter. We develop a CNN and BiLSTM deep learning module with an attention mechanism to extract and analyze data for precise workload prediction. Alibaba cluster workload is used in our trials to verify our model. According to experimental results, SG-CBA outperforms BTH-ARIMA, LSTNet, OCRO-MLNN, RNN, GRU, LSTM, and BiLSTM in workload prediction across a number of assessment metrics.

c) ALEDAR: An Attentions-based Encoder-Decoder and Autoregressive model for workload Forecasting of Cloud Data Center:

Effective workload predictions helps guide cloud data center resource scheduling. Multi-data centres offer more computing services and have more complex architectures than single data centres. The old load forecasting techniques require human parameter setting, while the new neural network approaches are insensitive to prediction scale and cannot capture long-term connections between input information. For these challenges, we propose a hybrid model called Attentions-based LSTM Encoder-Decoder network and Autoregressive model (ALEDAR) that uses neural network and statistical learning methods to analyze the linear and nonlinear load sequence over time in a multi-cloud data center. ALEDAR

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

employs a dual attention-based Encoder-Decoder architecture to recover historical workload data correlations and minimize long-range data scale-induced prediction impact degradation. The output layer is a three-layer perceptron. ALEDAR also uses an autoregressive module to capture the load sequence's linear trend and eliminate input and output scale insensitivity. Our adaptive technique improves host workload prediction in single- and multi-cloud data centers, according to experiments. The proposed technique outperforms state-of-the-art baselines by 9.7%–34.2% on real-world data sets.

d) Three-Way Ensemble Prediction for Workload in the Data Center:

A key component of cloud computing is accurate data center workload prediction, which is especially useful for increasing resource efficiency and lowering energy usage. However, it is difficult to get precise findings in cloud resource management due to the workload's quasi-volatility. In order to increase forecast accuracy, this research initially proposes a three-way ensemble prediction for workload in the data center. Additionally, we used a simulated annealing approach to determine the ideal threshold for dividing the workload after first defining it as the stable period, the volatility period, and the jitter period. In order to further increase the prediction accuracy, we then assigned several prediction models based on workload characteristics and a priori error prediction in accordance with the fundamental concept of the three-way decision. Lastly, TWD-RCPM improves workload forecast accuracy by 69.0%, 68.6%, and 72.6%, respectively, when compared to ARIMA, NN, and DMASVR-3WD using the CPU load monitoring logs from the Google cluster trace.

e) COSCO2: AI-augmented evolutionary algorithm based workload prediction framework for sustainable cloud data centers:

In the cloud data center, workload prediction is essential to preserving resource flexibility and scalability. However, due to noise, redundancy, and poor workload forecast accuracy in cloud data centers, the accuracy of workload prediction is extremely low. For sustainable cloud data centers, a tree hierarchical deep convolutional neural network (T-CNN) optimized using the sheep flock optimization technique is suggested in this article. First, the kernel correlation approach is used to preprocess the historical data from the cloud data center. In a dynamic cloud context, the suggested T-CNN technique is employed for workload prediction. The sheep flock optimization approach is used to optimize the weight parameters of the T-CNN model. The suggested COSCO2 approach lowers excessive power usage in cloud data centers and properly forecasts future workload. Two benchmark datasets are used to assess the suggested method: (i) NASA and (ii) Saskatchewan HTTP traces. This model is simulated using a Java tool, and the parameters are computed. According to the simulation, the suggested method achieves 20.64%, 32.95%, 12.05%, 32.65%, and 26.54% high accuracy and 27.4%, 26%, 23.7%, 34.7%, and 36.5% lower energy consumption for validating the NASA dataset, and 20.75%, 19.06%, 29.09%, 23.8%, 30.72%, and 33.74% lower energy consumption for validating the Saskatchewan HTTP traces dataset.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

3. METHODOLOGY

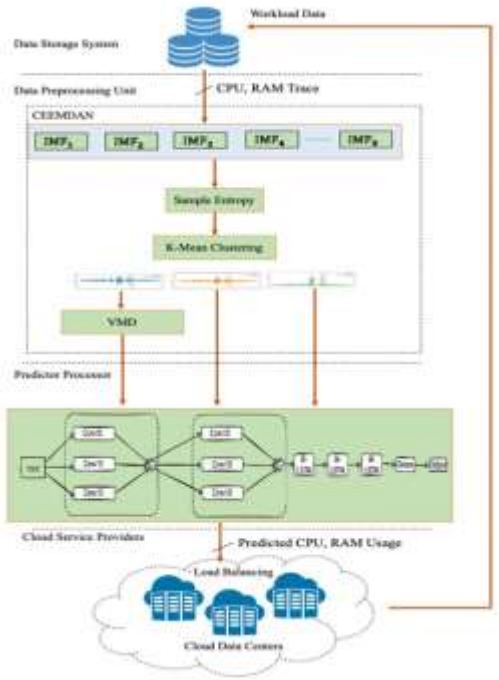
i) Proposed Work:

To enhance cloud data center workload prediction, the CVCBM design incorporates Bi-LSTM with a small Bidirectional Gated Recurrent Unit (BiGRU). While Sample Entropy (SE) and K-Means clustering identify high workload patterns for training, the two-step decomposition technique, which consists of CEEMDAN and VMD, preprocesses workload data to denoise signals and significant information. By training short-term events and long-term CPU use patterns, the bi-LSTM2BiGRU architecture can better capture all the multi-scale features of time dependence and forecast results more accurately than both the deep learning model and the traditional single-model.

Flask is used in the implementation of the improved model to provide an interactive interface, real-time workload prediction, and feasible deployment. Because customers may submit test files and get quick CPU use estimations, it performs well in dynamic cloud environments. The BiGRU layers are lightweight, decrease overfitting, enhance training performance, and minimize computation costs. The hybrid architecture balances time-critical information in both directions of the input series. For proactive resource allocation and energy-sensitive operations in large cloud data centers, this technology is scalable, robust, and effective.

ii) System Architecture:

Data storage, preprocessing, and prediction are the three main components of the extended workload prediction model's system architecture. CPU and RAM traces are gathered and kept in the data storage system. The preprocessing unit divides the workload raw data into Intrinsic Mode Functions (IMFs) with the help of CEEMDAN, then picks out the complex workload data with Sample Entropy, followed by classifying the workload data with K-Means as low, medium and high. In order to improve feature extraction, VMD is also used to extract high-frequency components. The hybrid Conv1D-BiLSTM-BiGRU network, which is capable of learning long-range dependencies and multi-scale temporal patterns, receives these characteristics via the prediction processor. Finally, the cloud service providers use the estimated CPU and RAM load to conduct the dynamic load balancing of the various cloud data centers, and thereby, permit the effective use of resources, less use of energy, and greater reliability of the services.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig1 proposed architecture

iii) Modules:

1. Data Collection Module

This module collects historical CPU and RAM workload traces from the Alibaba Cloud dataset. The collected workload data represents real cloud resource utilization patterns and serves as the input for preprocessing and prediction.

2. Data Preprocessing Module

This module preprocesses workload signals using CEEMDAN, Sample Entropy, and VMD techniques. It removes noise, decomposes workload signals into meaningful components, and extracts refined features for accurate prediction.

3. Workload Clustering Module

This module applies the K-Means clustering algorithm to categorize workloads into low, medium, and high workload groups. High workload data is prioritized for training to improve prediction reliability under critical resource conditions.

4. Hybrid Deep Learning Prediction Module

This module implements the hybrid Conv1D–BiLSTM–BiGRU architecture for workload forecasting. Conv1D extracts multi-scale temporal features, while BiLSTM and BiGRU capture short-term and long-term workload dependencies efficiently.

5. Evaluation Module

This module evaluates prediction performance using metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). It compares predicted workload values with actual values to measure model accuracy and stability.

6. Deployment Module

This module deploys the trained prediction model using the Flask web framework. It provides a user-friendly interface for uploading workload data and viewing real-time prediction results and performance outputs.

iv) Algorithms:

1. Support Vector Machine (SVM):

SVM is another common machine learning model that is utilized in regression to forecast workload. It is designed to identify the most appropriate hyperplane using the workload data. SVM performs appallingly in this case, as the graph illustrates, with extremely high MAE and MSE, suggesting that it is useless for managing the highly variable and high-dimensional nonlinear patterns of cloud workload.

2. Proposed CVCBM (Conv1D + Bi-LSTM + BiGRU):

Conv1D, which detects local temporal characteristics, Bi-LSTM, which detects long-term dependencies, and biGRU, which detects temporal features in a lightweight way, are all included into the suggested CVCBM hybrid model. It was designed to record multi-scale loads in cloud data centers. The graph indicates that the model is more accurate and reliable

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

than the conventional models since it has low MAE and MSE.

3. Extension CVCBM + BiGRU:

This is an extended version of the CVCBM model that successfully extracts temporal patterns while reducing computing complexity by stacking lightweight BiGRU layers. It is somewhat better than the original CVCBM in terms of the MAE and MSE, indicating that it is more efficient and predictive.

4. EXPERIMENTAL RESULTS

The experiment's results validate the advantages of the hybrid deep learning architecture and two-stage decomposition offered for improving workload forecast accuracy in cloud data centers. When CEEMDAN and VMD were used in the preprocessing pipeline, noise was much decreased, high-frequency and low-frequency workload components were easier to separate, and cleaner, more relevant input features were obtained. In order to guarantee that the high-impact workload segments were prioritized and allow the model to learn the crucial fluctuations that are relevant in an even better fashion, the composite design that incorporated Sample Entropy-based IMF selection and K-Means clustering was important.

By adding lightweight BiGRU layers to the Bi-LSTM and achieving multi-scale temporal learning with lower computing costs, the hybrid CVCBM model was also improved in accordance with the concept of extension. Compared to the baseline CVCBM and conventional machine learning methods like SVM, the long model was more accurate. According to the performance graph, the expanded CVCBM + BiGRU model has an MAE of 0.0074, which is much better than SVM (MAE 0.092) and higher than the original CVCBM (MAE 0.0082). It was discovered that the combination of Conv1D, Bi-LSTM, and Bi-GRU could anticipate the workload that would be used in the cloud resources optimization job by modeling both short-term and long-term temporal dependencies in a stable and reliable manner.

Overall, the results of the experiments show that the hybrid architecture significantly improves prediction accuracy, computing efficiency, and the potential for real-time deployment using Flask for dynamic workload forecasting in large-scale cloud data centers.

Accuracy: A test's accuracy is its capacity to distinguish healthy from ill cases. Find the percentage of instances with genuine positives and negatives to assess test accuracy.

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Classification accuracy or positive cases constitute precision. The formula for accuracy is:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's recall measures its ability to recognize all appropriate machine learning class instances. The ratio of accurately predicted positive

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

observations to total positives indicates a model's class instance detection skill.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision ranks quality. It considers the number and order of relevant ideas. Calculating MAP at K uses the arithmetic mean of each user or query's Average Precision (AP).

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$
$$AP_k = \text{the AP of class } k$$
$$n = \text{the number of classes}$$

F1-Score: A high F1 score suggests an accurate machine learning model. Integrating recall and precision improves model correctness. Accuracy measures how often a model predicts a dataset correctly.

$$F1 = 2 \cdot \frac{(Recall \cdot \text{Pr}\, e\, cision)}{(Recall + \text{Pr}\, e\, cision)}$$

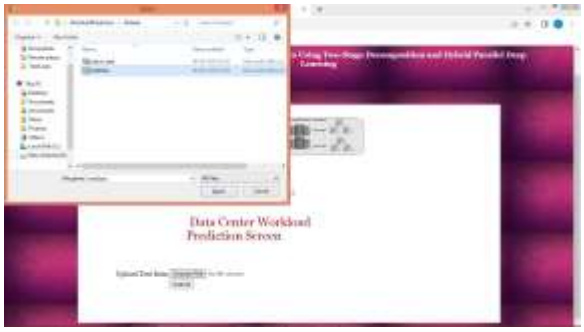


Fig2 upload image

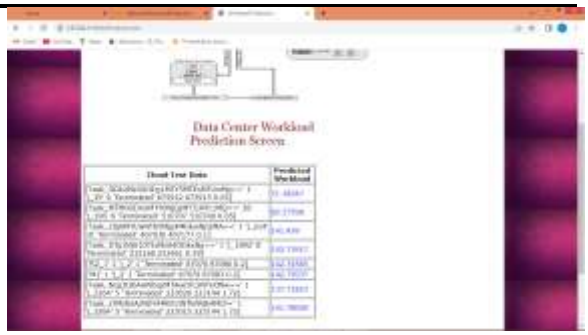


Fig2 Results

5. CONCLUSION

The proposed system successfully improves cloud workload prediction by integrating two-stage signal decomposition with a hybrid deep learning framework. CEEMDAN, Sample Entropy, and VMD effectively preprocess and refine noisy workload data, while the Conv1D–BiLSTM–BiGRU architecture accurately captures multi-scale temporal dependencies and workload variations.

Experimental evaluation using Alibaba Cloud workload traces demonstrates improved prediction accuracy with reduced MAE and MSE compared to traditional models. The proposed approach supports proactive resource provisioning, load balancing, energy-efficient cloud management, and real-time workload forecasting, making it suitable for modern large-scale cloud data center environments.

6. FUTURE SCOPE

The proposed cloud workload prediction framework can be further enhanced by extending it to multi-cloud and distributed cloud environments for large-scale resource management and intelligent workload balancing. Future improvements may include integrating advanced deep learning techniques such

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

as attention mechanisms, Transformer models, and reinforcement learning to further improve prediction accuracy and adaptive decision-making capabilities. The system can also be combined with automated resource provisioning, virtual machine migration, and container orchestration platforms such as Kubernetes to enable fully autonomous cloud infrastructure management.

In addition, future work can incorporate multiple cloud performance metrics including network traffic, storage utilization, and energy consumption for comprehensive workload analysis and optimization. GPU-based acceleration and edge-cloud integration may also be explored to reduce prediction latency and improve scalability for real-time applications. Deploying the framework on live industrial cloud platforms can further validate its effectiveness in practical cloud computing environments.

REFERENCES

[1] H. Yuan, J. Bi, and M. Zhou, ‘‘Multiqueue scheduling of heterogeneous tasks with bounded response time in hybrid green IaaS clouds,’’ IEEE Trans. Ind. Informat., vol. 15, no. 10, pp. 5404–5412, Oct. 2019.

[2] (2020). Cisco Global Cloud Index. [Online]. Available: https://www.cisco. com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/ white-paper-c11-738085.html

[3] L. A. Barroso and U. Hlzle, The Datacenter As a Computer: An Introduction To the Design of Warehouse-Scale Machines. San Rafael, CA, USA: Morgan & Claypool, 2009.

[4] J. Bi, H. Yuan, W. Tan, M. Zhou, Y. Fan, J. Zhang, and J. Li, ‘‘Applicationaware dynamic fine-grained resource provisioning in a virtualized cloud data center,’’ IEEE Trans. Autom. Sci. Eng., vol. 14, no. 2, pp. 1172–1184, Apr. 2017.

[5] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ‘‘Empirical evaluation of gated recurrent neural networks on sequence modeling,’’ 2014, arXiv:1412.3555.

[7] J. Bi, H. Yuan, K. Zhang, and M. Zhou, ‘‘Energy-minimized partial computation offloading for delay-sensitive applications in heterogeneous edge networks,’’ IEEE Trans. Emerg. Topics Comput., vol. 10, no. 4, pp. 1941–1954, Oct. 2022.

[8] H. Yuan, J. Bi, and M. Zhou, ‘‘Geography-aware task scheduling for profit maximization in distributed green data centers,’’ IEEE Trans. Cloud Comput., vol. 10, no. 3, pp. 1864–1874, Jul. 2022.

[9] S. Li, Y. Wang, X. Qiu, D. Wang, and L. Wang, ‘‘A workload predictionbased multi-VM provisioning mechanism in cloud computing,’’ in Proc. 15th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS), Sep. 2013, pp. 1–6.

[10] M. Barati and S. Sharifian, ‘‘A hybrid heuristic-based tuned support vector regression model for cloud load prediction,’’ J. Supercomput., vol. 71, no. 11, pp. 4235–4259, Nov. 2015.

[11] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, ‘‘Workload prediction using ARIMA model and its impact on cloud applications’ QoS,’’ IEEE

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Trans. Cloud Comput., vol. 3, no. 4, pp. 449–458, Oct. 2015, doi: 10.1109/TCC.2014.2350475.

[12] Q. Sun, Z. Tan, and X. Zhou, ‘‘Workload prediction of cloud computing based on SVM and BP neural networks,’’ J. Intell. Fuzzy Syst., vol. 39, no. 3, pp. 2861–2867, Oct. 2020, doi: 10.3233/jifs-191266.

[13] Y. Bao, T. Xiong, and Z. Hu, ‘‘Multi-step-ahead time series prediction using multiple-output support vector regression,’’ Neurocomputing, vol. 129, pp. 482–493, Apr. 2014.

[14] Y. Lu, J. Panneerselvam, L. Liu, and Y. Wu, ‘‘RVLBPNN: A workload forecasting model for smart cloud computing,’’ Sci. Program., vol. 2016, pp. 1–9, Nov. 2016.

[15] M. Amiri and L. Mohammad-Khanli, ‘‘Survey on prediction models of applications for resources provisioning in cloud,’’ J. Netw. Comput. Appl., vol. 82, pp. 93–113, Mar. 2017, doi: 10.1016/j.jnca.2017.01.016.

AUTHOR’S PROFILE

Mr. K. Uday Kiran is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Bapatla Engineering College, Bapatla. His research interests include Machine Learning Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. T. DREEM is an MCA Student in the Department of Computer Application at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He has Completed Degree in MPCS from Sri vasavi kanyaka parameswari arts. science & commerce markapur college , Prakasam district.