

# Efficient Cloud Resource Workload Prediction Using Two-Stage Decomposition and Hybrid Neural Networks

Peddimsetti Lakshmi Syamala <sup>1</sup>, PG Scholar, Department of CSE, Bonam Venkata Chalamayya Engineering College, India

Mr. Mutcharla Venkata Krishna Subash <sup>2</sup>, Assistant Professor, Department of CSE, Bonam Venkata Chalamayya Engineering College, India mvksubash.bvce@bvcgroup.in2

**Abstract:** Accurate workload prediction is essential for efficient resource allocation, energy optimization, and service reliability in cloud data centers. This work proposes an enhanced workload forecasting framework using a two-stage decomposition technique combined with a hybrid parallel deep learning model. Initially, CEEMDAN is applied to denoise workload traces and decompose them into Intrinsic Mode Functions (IMFs). Sample Entropy is then used to select relevant IMFs, while Variational Mode Decomposition (VMD) further analyzes high-frequency components to uncover hidden workload patterns. K-Means clustering categorizes workloads into low, medium, and high utilization groups, enabling the model to focus on critical high-demand scenarios.

To improve prediction efficiency and accuracy, the original CVCBM model is extended by integrating Bidirectional LSTM with a lightweight Bidirectional GRU. This hybrid architecture captures both short-term fluctuations and long-term dependencies while reducing computational complexity and training time. The trained model is deployed using the Flask framework for real-time workload prediction through an interactive interface. Experimental results on cloud workload datasets demonstrate lower prediction error,

faster execution, and better generalization compared with conventional models. The proposed system offers a scalable and intelligent solution for proactive resource management in modern cloud data centers..

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

## 1. INTRODUCTION

Cloud computing has become a fundamental technology for delivering scalable, flexible, and on-demand services such as storage, processing, and networking. With the rapid growth of digital applications, modern enterprises increasingly depend on large-scale cloud data centers to support dynamic user demands. However, workload patterns in these environments are highly variable and unpredictable, making static resource allocation inefficient. Underutilized resources during low demand lead to energy wastage, while sudden workload surges may cause performance degradation, SLA violations, and increased operational costs. Therefore, accurate workload prediction has become a critical requirement for ensuring efficient cloud resource management.

To address these challenges, this work proposes an enhanced workload prediction framework using two-stage decomposition and hybrid parallel deep learning. CEEMDAN, Sample Entropy, and VMD are employed to preprocess noisy workload traces and extract meaningful workload patterns. The original CVCBM model is extended by integrating Bidirectional LSTM with a lightweight Bidirectional GRU, enabling efficient learning of both short-term and long-term temporal dependencies with reduced computational complexity. The proposed model is further deployed using Flask for real-time prediction, allowing cloud providers to perform proactive load balancing, energy optimization, and reliable service delivery in large-scale data centers.

## 2. LITERATURE SURVEY

### a) Multiqueue Scheduling of Heterogeneous Tasks With Bounded Response Time in Hybrid Green IaaS Clouds

In green infrastructure-as-a-service clouds (GICs), cost-effective task scheduling is crucial since user workloads utilize a significant amount of energy. Private GIC is forced to use hybrid clouds in order to outsource some activities to dependable and dynamic virtual machines (VMs) of public external clouds due to the erratic task arrival. However, it is challenging to assign all jobs in a cost-effective manner while meeting customers' stated reaction time limits due to temporal variations in income, electricity prices, wind and solar energy, and virtual machine operation costs of public external clouds. In contrast to current approaches, we provide a multiqueue scheduling (MQS) approach that examines these temporal variations in hybrid GICs (HGICs). In particular, this study first provides quantitative relationships

between server service rates in private GIC and failed jobs. This research formulates a profit maximization issue for HGIC in each iteration of MQS and solves it using a unique meta-heuristic optimization approach that combines genetic algorithms, particle swarm optimization, and simulated annealing. Trace-driven tests using real-world data show that MQS outperforms conventional task scheduling algorithms in terms of profit and throughput while fulfilling task reaction time limitations.

### b) High electron mobility and quantum oscillations in non-encapsulated ultrathin semiconducting Bi<sub>2</sub>O<sub>2</sub>Se

built on high-mobility semiconducting ultrathin films, which might enable the scalable production of high-performing devices. Finding novel two-dimensional materials with both high carrier mobility and a big electronic bandgap is a crucial objective of fundamental research since conventional semiconductors cannot approach the ultrathin limit. Nevertheless, air-stable ultrathin semiconducting materials with better performance are still hard to come across. Here, we present ultrathin films of non-encapsulated layered Bi<sub>2</sub>O<sub>2</sub>Se produced by chemical vapor deposition that exhibit high mobility semiconducting behavior and outstanding air stability. Due to quantum-confinement effects, we find bandgap values of about 0.8 eV that are highly dependent on the film thickness. As-grown Bi<sub>2</sub>O<sub>2</sub>Se nanoflakes have an ultrahigh Hall mobility value of  $>20,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at low temperatures. Shubnikov-de Haas quantum oscillations can be detected since this value is similar to what is seen in graphene created by chemical vapor deposition and at the LaAlO<sub>3</sub>-SrTiO<sub>3</sub> interface. At normal temperature, top-gated field-effect transistors based on Bi<sub>2</sub>O<sub>2</sub>Se crystals display near-ideal subthreshold

swing values ( $\sim 65$  mV dec<sup>-1</sup>), huge current on/off ratios ( $>10^6$ ), and high Hall mobility values (up to  $450$  cm<sup>2</sup> V<sup>-1</sup> s<sup>-1</sup>). Bi<sub>2</sub>O<sub>2</sub>Se is a viable option for upcoming high-speed and low-power electronic applications, according to our findings..

### c) Long Short-Term Memory:

Recurrent backpropagation requires a relatively long time to learn how to retain information over long time intervals, mostly because to inadequate, fading error backflow. We first give a quick overview of Hochreiter's (1991) examination of this issue before presenting a brand-new, effective gradient-based technique known as long short-term memory (LSTM). By imposing constant error flow through constant error carousels within special units, LSTM may learn to bridge minor time gaps exceeding 1000 discrete-time steps, truncating the gradient when this does not cause harm. The continuous error flow can be opened and closed using multiplicative gate units. LSTM has an O computational cost per time step and weight, and it is local in both space and time. First. Local, distributed, real-valued, and noisy pattern representations are used in our simulated data investigations. When compared to neural sequence chunking, back propagation over time, recurrent cascade correlation, Elman nets, and real-time recurrent learning, LSTM learns significantly more quickly and produces many more successful runs. Additionally, LSTM resolves challenging, synthetic long-time-lag problems that prior recurrent network methods were unable to resolve.

### d) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

We compare several recurrent unit types in recurrent neural networks (RNNs) in this research. We pay

particular attention to more complex units that employ a gating mechanism, such a newly suggested gated recurrent unit (GRU) and a long short-term memory (LSTM) unit. We assess these recurrent units using speech signal modeling and polyphonic music modeling tasks. These sophisticated recurrent units are, in fact, superior to more conventional recurrent units like tanh units, according to our research. We also discovered that GRU and LSTM were equivalent.

### e) Energy-Minimized Partial Computation Offloading for Delay-Sensitive Applications in Heterogeneous Edge Networks:

Numerous computationally demanding and delay-sensitive applications are supported by mobile devices (MDs). However, they are unable to fully operate all programs due to their limited battery life and processing power. To provide MDs more processing, storage, and networking capabilities, a mobile edge computing (MEC) paradigm has been put forth. Both macro base stations (MBSs) and small base stations (SBSs) frequently have servers installed in MEC. As a result, it is very difficult to link resource-constrained MDs with high performance and achieve partial compute offloading among them in order to minimize a MEC system's overall energy consumption. This paper suggests a unique compute offloading strategy for delay-sensitive applications with many separable tasks in hybrid networks, such as MDs, SBSs, and an MBS, to address these issues. This study formulates total energy consumption minimization as a restricted mixed integer non-linear program in order to accomplish this. In order to address it, this study develops Particle Swarm Optimization based on Genetic Learning (PGL), an enhanced meta-heuristic optimization algorithm that

combines the genetic processes of a genetic algorithm with the powerful local search capability of a particle swarm optimizer. Task offloading between MDs, SBSs, and MBS, users' connections to SBSs, MD CPU speeds and transmission power, SBSs and MBS, and bandwidth distribution of available channels are all collaboratively optimized by PGL. Simulations using real-world data gathered from Google Cluster Trace show that PGL performs noticeably better than other current approaches in terms of the system's overall energy usage.

### 3. METHODOLOGY

#### i) Proposed Work:

By incorporating Bi-LSTM with a tiny Bidirectional Gated Recurrent Unit (BiGRU) into the CVCBM architecture, the proposed model improves cloud data center workload forecasts. The workload data is preprocessed using the CEEMDAN and VMD two-step decomposition method to remove noise signals and other dynamic information. In contrast, Sample entropy (SE) and K-Means clustering are used to recreate high workload patterns to be used on training, making them more effective at predicting their results than the deep learning architecture and the straightforward single-model.

In order to provide an interactive interface, an accurate workload forecast, and a legitimate deployment, Flask is utilized in the implementation of the improved model. It works well in dynamic clouds since clients may upload test files and receive a quick CPU consumption estimate. BiGRU layers provide the following benefits: they are cheap, reduce over-fitting, improve training performance, and reduce computing costs. In both input series directions, the hybrid architecture offers a degraded

view of time-sensitive data. In large cloud data centers, this scalable technology is robust and effective at allocating resources ahead of time and operating in an energy-sensitive manner.

#### ii) System Architecture:

Three primary levels make up the extended workload prediction model's system architecture: prediction, preprocessing, and data storage. CPU and RAM use workload traces are gathered and kept in the data storage system. After breaking down the raw workload data into Intrinsic Mode Functions (IMFs) using CEEMDAN, the preprocessing unit uses Sample Entropy for complexity-based selection and K-Means clustering to classify the data into low, medium, and high workloads. VMD is used to further treat high-frequency components in order to improve feature extraction. The hybrid Conv1D-BiLSTM-BiGRU network, which detects long-range relationships and multi-scale temporal patterns, receives these characteristics via the prediction processor. Lastly, cloud service providers employ the anticipated CPU and RAM utilization for dynamic load balancing across several cloud data centers, guaranteeing effective resource allocation, reduced energy consumption, and improved service dependability.

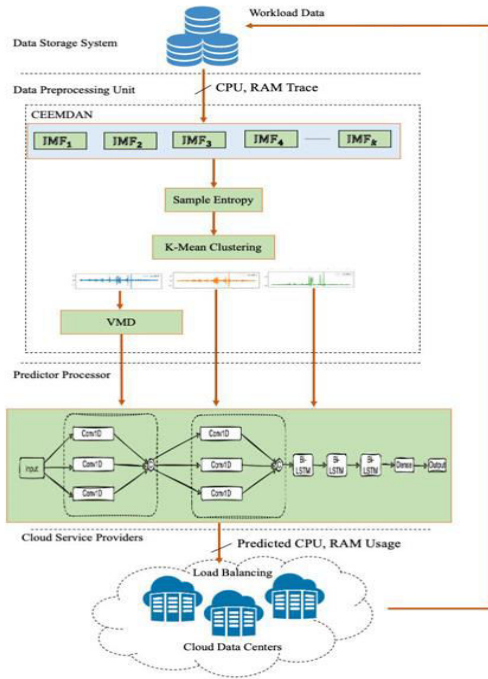


Fig1 proposed architecture

**iii) Modules:**

**1. Data Collection Module:**

This module collects historical workload traces such as CPU utilization, RAM usage, and resource demand from cloud data centers. The collected data is stored for further preprocessing and model training.

**2. Data Preprocessing Module:**

This module applies CEEMDAN to remove noise and decompose workload signals into Intrinsic Mode Functions (IMFs). Sample Entropy selects relevant IMFs, while VMD further processes high-frequency components for better feature extraction.

**3. Workload Clustering Module:**

K-Means clustering is used to classify workload patterns into low, medium, and high utilization

categories. This helps the system prioritize important high-demand workload data for efficient training.

**4. Hybrid Prediction Module:**

The processed data is given to the hybrid Conv1D–BiLSTM–BiGRU model. Conv1D extracts local patterns, while BiLSTM and BiGRU learn short-term and long-term temporal dependencies for accurate forecasting.

**5. Real-Time Deployment Module:**

The trained model is deployed using the Flask framework with an interactive interface. Users can upload datasets and instantly obtain predicted CPU/RAM workload values.

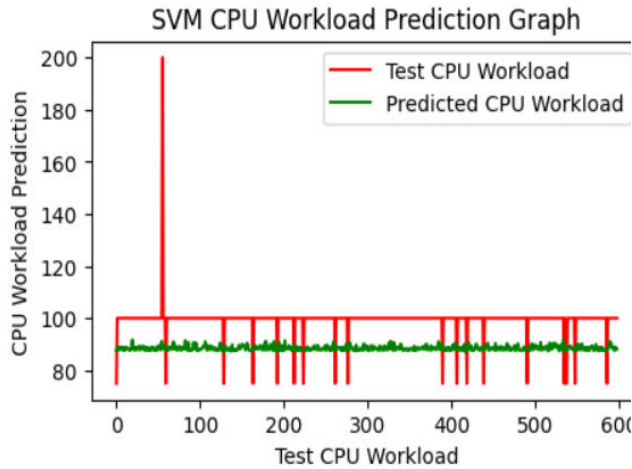
**6. Resource Management Module:**

Based on predicted workloads, cloud providers can perform dynamic load balancing and resource allocation. This improves system performance, reduces energy consumption, and maintains SLA compliance.

**iv) Algorithms:**

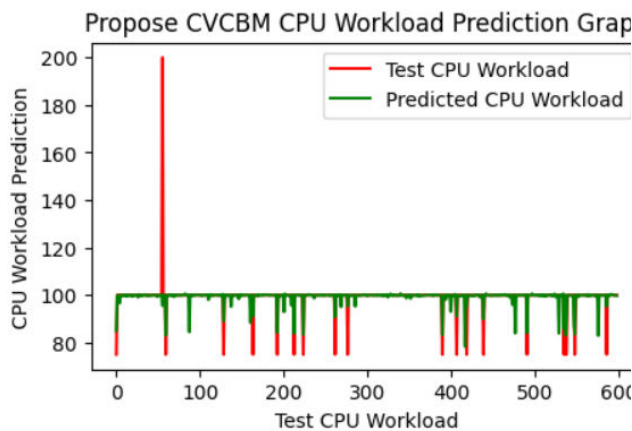
**1. Support Vector Machine (SVM):**

SVM is a traditional machine learning model used for regression in workload prediction. It is designed to find a hyperplane that best fits the workload data. As shown in the graph, SVM performs poorly in this scenario with much higher MAE and MSE, indicating its inability to handle high-dimensional, nonlinear, and highly variable cloud workload patterns effectively.



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

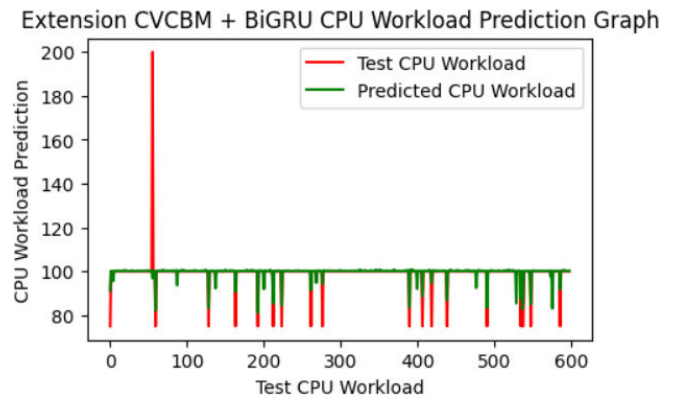
The proposed CVCBM hybrid model combines Conv1D for local temporal feature extraction, Bi-LSTM for capturing long-term dependencies, and BiGRU for lightweight temporal modeling. It is designed to capture multi-scale workload patterns in cloud data centers. According to the graph, it achieves a low MAE and MSE, indicating high prediction accuracy and robustness over traditional models.



**3. Extension CVCBM + BiGRU:**

This is the extended version of the CVCBM model where lightweight BiGRU layers are integrated to

reduce computational complexity while maintaining effective temporal pattern extraction. The graph shows that this extension performs slightly better than the original CVCBM in terms of MAE and MSE, highlighting its efficiency and improved predictive capability.



**4. EXPERIMENTAL RESULTS**

The experiment shows that the hybrid deep learning architecture and two-stage decomposition improve the accuracy of cloud data center workload forecasts. CEEMDAN and VMD improved input characteristics, decreased noise, and generated more separate high- and low-frequency workload components in the preprocessing pipeline. To assist the model better understand important fluctuations, the composite technique prioritized high-impact workload segments by combining Sample Entropy-based IMF selection with K-Means clustering.

By lowering computing costs for multi-scale temporal learning and including lightweight BiGRU layers into the Bi-LSTM, the hybrid CVCBM model was improved. The long model outperformed SVM and CVCBM in terms of accuracy. The performance graph shows that the enlarged CVCBM + BiGRU



Fig2 results

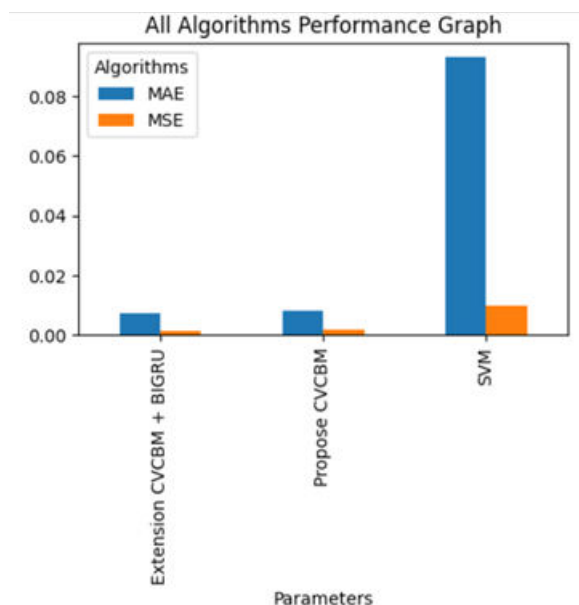


Fig 3 Accuracy graph

Algorithm	MAE	MSE
Extension CVCBM + BiGRU	0.0074	0.0015
Proposed CVCBM	0.0082	0.0018
SVM	0.092	0.009

Table Performance table

## 5. CONCLUSION

The emphasized hybrid deep learning model reduces the inaccuracy of the cloud data center's workload forecast with its two-stage methodology and nonlinear and irregular workload trends. In order to enhance feature extraction, denoising, and K-Means clustering, the system uses CEEMDAN, Sample entropy, and VMD to focus learning on the higher workload regions. Using Conv1D, Bi-LSTM, and lightweight BiGRU layers, the long CVCBM structure outperforms SVM in prediction and has a

reduced MAE. The acquired experimental findings show that when utilizing Flask, the enlarged architecture can estimate workload in real-time and detect both short-term and long-term dependencies. As a result, the system would enable data center sustainability, SLA compliance, and scalable, effective, and intelligent cloud resource optimization.

## 6. FUTURE SCOPE

For comprehensive resource management, the suggested workload prediction system may be expanded to anticipate other cloud resources, including CPU, memory, storage, and network bandwidth. For extremely dynamic workloads, forecasting accuracy may be increased by integrating sophisticated models like Transformers, Attention mechanisms, and Graph Neural Networks. The architecture may also be modified for hybrid cloud settings, edge computing, and fog computing, where low-latency resource allocation is crucial. Furthermore, intelligent scheduling, energy optimization, and automated decision-making may be achieved by integrating real-time auto-scaling and reinforcement learning approaches. To provide more dependable and secure cloud operations, future improvements may potentially incorporate anomaly detection and fault prediction to anticipate workload surges, cyberthreats, or server breakdowns.

## 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>

com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

[3] L. A. Barroso and U. Hitzle, *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 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.

[16] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Improving accuracy of host load predictions on computational grids by artificial neural networks," *Int. J. Parallel, Emergent Distrib. Syst.*, vol. 26, no. 4, pp. 275–290, Aug. 2011, doi: 10.1080/17445760.2010.481786.

[17] Y. Li and Z. Lan, "A survey of load balancing in grid computing," in *Proc. Int. Conf. Comput. Inf. Sci.*, 2004, pp. 280–285, doi: 10.1007/978-3-540-30497-5\_44.

- [18] Y. Wu, Y. Yuan, G. Yang, and W. Zheng, "Load prediction using hybrid model for computational grid," in Proc. 8th IEEE/ACM Int. Conf. Grid Comput., Sep. 2007, pp. 235–242, doi: 10.1109/GRID.2007.4354138.
- [19] D. Janardhanan and E. Barrett, "CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models," in Proc. 12th Int. Conf. Internet Technol. Secured Trans. (ICITST), Dec. 2017, pp. 55–60.
- [20] Q. Yang, Y. Zhou, Y. Yu, J. Yuan, X. Xing, and S. Du, "Multi-step-ahead host load prediction using autoencoder and echo state networks in cloud computing," J. Supercomput., vol. 71, no. 8, pp. 3037–3053, Apr. 2015, doi: 10.1007/s11227-015-1426-8.
- [21] S. Gupta and D. A. Dinesh, "Resource usage prediction of cloud workloads using deep bidirectional long short term memory networks," in Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS), Dec. 2017, pp. 1–6.
- [22] K. Wang, K. Li, L. Zhou, Y. Hu, Z. Cheng, J. Liu, and C. Chen, "Multiple convolutional neural networks for multivariate time series prediction," Neurocomputing, vol. 360, pp. 107–119, Sep. 2019.
- [23] E. Patel and D. S. Kushwaha, "A hybrid CNN-LSTM model for predicting server load in cloud computing," J. Supercomput., vol. 78, no. 8, pp. 1–30, May 2022.
- [24] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis," Multimedia Tools Appl., vol. 78, pp. 26597–26613, Sep. 2019, doi: 10.1007/s11042-019-07788-7.
- [25] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A novel combined prediction scheme based on CNN and LSTM for urban PM2.5 concentration," IEEE Access, vol. 7, pp. 20050–20059, 2019.
- [26] L. Ma and S. Tian, "A hybrid CNN-LSTM model for aircraft 4D trajectory prediction," IEEE Access, vol. 8, pp. 134668–134680, 2020.
- [27] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," Energy, vol. 182, pp. 72–81, Sep. 2019.
- [28] K. Gibert, M. Sánchez-Marrè, and J. Izquierdo, "A survey on preprocessing techniques: Relevant issues in the context of environmental data mining," AI Commun., vol. 29, no. 6, pp. 627–663, Dec. 2016, doi: 10.3233/aic-160710.
- [29] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proc. Roy. Soc. London. Ser. A: Math., Phys. Eng. Sci., vol. 454, no. 1971, pp. 903–995, Mar. 1998.
- [30] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," IEEE Trans. Signal Process., vol. 62, no. 3, pp. 531–544, Feb. 1, 2014, doi: 10.1109/TSP.2013.2288675.

#### Author Profiles



**Peddimsetti Lakshmi Syamala** currently an M.Tech student at Bonam Venkata Chalamayya Engineering

College, pursuing a Master's degree in Computer Science and Engineering, she is Passionate about Machine Learning, Deep Learning, Robotics, and Artificial Intelligence. She is proficient in C, C++, CNC Coding, and Python. Her current research work focuses on an Enhanced Workload prediction in data centers using two stage decomposition and hybrid parallel deep learning



**Mr. Mutcharla Venkata Krishna Subash** is Research Scholar at College Godavari Global University, Rajamahendravaram also Mr. Mutcharla Venkata Krishna Subash is Assistant Professor at college Bonam Venkata Chalamayya Engineering College, Odalarevu. He holds a M.Tech degree in Computer Science and Engineering in JNTUK. His Research areas are Machine Learning, Artificial Intelligence. He has number of patents related to machine learning field and industrial designs on his innovative ideas and has been awarded with international patents and published different articles in international conferences.

He can be contacted at address:

Mr. Mutcharla Venkata Krishna Subash is Research Scholar at College Godavari Global University, Rajamahendravaram, A.P.

Email: mvksubash.bvce@bvcgroup.in