International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

An Intelligent Hybrid CEEMDAN-CNN-BiLSTM-GRU Framework for Dynamic Cloud Workload Prediction and Resource Optimization

Kondapalli Likhitha[1], S. Venkateswara Rao[2]

[1]M.Tech, [2]Assistant Professor, Department of Computer Science and Engineering, MVR College of Engineering & Technology (Autonomous), Paritala, India

***Abstract*-- Cloud computing environments generate highly dynamic and non-linear workload patterns due to continuous variations in user requests, virtual machine operations, and distributed services. Accurate workload prediction is essential for efficient resource allocation, reducing energy consumption, improving virtual machine utilization, and maintaining service-level agreements in cloud data centers. However, traditional statistical methods and standalone deep learning models often fail to achieve high prediction accuracy because they cannot effectively capture complex temporal dependencies and fluctuating workload behaviors.**

To address these challenges, this paper proposes an intelligent hybrid deep learning framework that combines Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Units (GRU) for dynamic cloud workload prediction and resource optimization. Initially, CEEMDAN is used to decompose complex workload signals into multiple intrinsic mode functions for noise reduction and improved signal stability. CNN layers are then utilized to extract significant workload features, while BiLSTM captures bidirectional temporal dependencies from historical workload sequences. Finally, GRU layers optimize sequential learning and reduce computational complexity during prediction.

The proposed system is implemented using Python, TensorFlow, Keras, and Flask to provide a practical real-time workload forecasting environment. Experimental evaluation is conducted using cloud workload datasets, and performance is analyzed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The obtained results demonstrate that the proposed hybrid framework achieves better prediction accuracy and resource optimization compared to conventional machine learning and single deep learning models. The developed framework provides an efficient and scalable solution for intelligent cloud workload management in modern cloud computing infrastructures.

***Keywords*-- Cloud Workload Prediction, Cloud Computing, CEEMDAN, Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), Gated Recurrent Unit (GRU), Deep Learning, Resource Optimization, Time-Series Forecasting, Hybrid Neural Network, Virtual Machine Management, Intelligent Resource Allocation**

I. Introduction

Cloud computing has transformed the modern computing landscape by providing scalable, on-demand, and cost-effective computing resources through the internet. It enables organizations and individuals to access computing services such as storage, processing power, networking, virtualization, and software platforms without maintaining dedicated physical infrastructure. Due to the rapid adoption of cloud-based services, applications such as e-commerce, healthcare systems, online education, multimedia streaming, artificial intelligence, Internet of Things (IoT), and big data analytics heavily depend on cloud computing infrastructures for continuous service availability and performance. As the number of users and cloud applications continues to increase, cloud data centers experience highly dynamic and unpredictable workload patterns, creating major challenges in resource allocation and workload management.

Efficient resource utilization is one of the most critical objectives in cloud computing environments. Cloud service providers must dynamically allocate computational resources such as CPU, memory, bandwidth, and virtual machines according to continuously changing workloads. Improper resource allocation can lead to several issues including resource underutilization, server overloading, increased operational costs, excessive energy consumption, reduced Quality of Service (QoS), and violations of Service Level Agreements (SLAs). Therefore, accurate cloud workload prediction has become an essential requirement for intelligent resource management and maintaining the stability of cloud infrastructures.

Cloud workload prediction refers to the process of forecasting future resource demands based on historical workload patterns and system behavior. Accurate prediction allows cloud systems to proactively allocate resources before workload congestion occurs, thereby improving system efficiency and reducing latency. However, cloud workload data is highly nonlinear, non-stationary, noisy, and time-dependent in nature. Workload patterns continuously fluctuate due to varying user requests, virtual machine migrations, distributed application execution, and real-time traffic variations.

These characteristics make workload forecasting a highly complex task.

Traditional statistical prediction approaches such as Moving Average, Linear Regression, Auto-Regressive Integrated Moving Average (ARIMA), and Exponential Smoothing have been widely used for time-series forecasting in cloud environments. Although these methods are computationally simple and effective for stable datasets, they often fail to capture complex nonlinear dependencies and temporal variations present in modern cloud workload data. As cloud infrastructures become more dynamic and large-scale, traditional approaches are no longer sufficient for achieving accurate and reliable workload prediction.

To overcome these limitations, machine learning and deep learning techniques have gained significant attention in cloud workload forecasting research. Machine learning models such as Support Vector Machines (SVM), Decision Trees, Random Forest, and Gradient Boosting have demonstrated improved predictive capabilities compared to conventional statistical methods. Furthermore, deep learning architectures including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) have shown remarkable performance in analyzing sequential and time-series data due to their ability to learn hidden patterns and temporal dependencies automatically.

Among deep learning models, recurrent neural networks such as LSTM and GRU are widely used for workload forecasting because they can effectively capture long-term temporal relationships from sequential cloud workload data. Similarly, CNN models are capable of extracting important local features and hidden workload characteristics from input sequences. Despite these advantages, standalone deep learning models still suffer from several limitations such as overfitting, high computational complexity, gradient vanishing problems, and insufficient feature extraction when handling highly fluctuating workload signals. Moreover, the presence of noise and irregular workload variations significantly affects the prediction performance of deep learning models.

Recently, hybrid deep learning and signal decomposition approaches have emerged as effective solutions for improving workload prediction accuracy. Signal decomposition techniques are particularly useful for separating complex workload signals into multiple stable components before feeding them into deep learning architectures. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an advanced signal decomposition technique capable of decomposing nonlinear and non-stationary time-series signals into several Intrinsic Mode Functions (IMFs).

This decomposition process helps reduce signal noise, stabilize workload fluctuations, and improve the learning capability of neural networks. By preprocessing workload data using CEEMDAN, deep learning models can more effectively identify meaningful temporal patterns and hidden workload behaviors.

Motivated by these challenges, this paper proposes an intelligent hybrid deep learning framework that integrates CEEMDAN, Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Units (GRU) for dynamic cloud workload prediction and resource optimization. In the proposed framework, CEEMDAN is initially applied to decompose cloud workload signals into multiple stable intrinsic mode functions. CNN layers are then utilized to extract significant local workload features from decomposed signals. The extracted features are further processed using BiLSTM layers to capture bidirectional temporal dependencies and historical workload relationships. Finally, GRU layers optimize sequential learning and reduce computational complexity while improving forecasting efficiency.

The proposed system is implemented using Python, TensorFlow, Keras, and Flask to provide a practical real-time workload prediction environment. The developed framework supports intelligent workload forecasting, predictive resource allocation, workload balancing, and efficient cloud infrastructure management. Experimental evaluation is conducted using cloud workload datasets, and system performance is analyzed using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and prediction accuracy. The obtained results demonstrate that the proposed hybrid framework achieves significantly better forecasting performance compared to traditional machine learning methods and standalone deep learning architectures.

The major contribution of this research lies in the development of a practical and scalable hybrid cloud workload prediction framework that combines advanced signal decomposition techniques with multiple deep learning architectures for intelligent resource optimization. The integration of CEEMDAN with CNN, BiLSTM, and GRU enables the proposed system to effectively handle nonlinear workload fluctuations, improve prediction stability, and enhance cloud resource utilization. The developed framework can support modern cloud infrastructures in achieving efficient workload management, reduced operational costs, energy optimization, and improved service reliability.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

II. Literature Review

Cloud workload prediction has emerged as a critical research area in cloud computing due to the rapid growth of virtualized infrastructures, distributed applications, and large-scale internet services.

Efficient workload forecasting helps cloud providers optimize resource utilization, improve virtual machine scheduling, reduce operational costs, minimize energy consumption, and maintain Service Level Agreements (SLAs). Over the years, researchers have proposed several statistical, machine learning, and deep learning approaches for predicting cloud workloads. However, accurately forecasting highly dynamic and nonlinear cloud workloads remains a challenging problem.

Initially, traditional statistical methods such as Linear Regression, Moving Average, Exponential Smoothing, and Auto-Regressive Integrated Moving Average (ARIMA) were widely used for workload forecasting. These methods were suitable for small-scale and stable datasets due to their low computational complexity and simple implementation. Among them, ARIMA became one of the most commonly used time-series forecasting techniques because it could model workload trends and sequential dependencies effectively. However, cloud workload data is highly nonlinear and non-stationary in nature, containing sudden spikes and irregular fluctuations caused by varying user requests and distributed service operations. Traditional statistical models fail to effectively capture these complex patterns, resulting in reduced prediction accuracy and poor adaptability in real cloud environments.

To overcome the limitations of statistical approaches, machine learning algorithms were introduced for cloud workload prediction. Techniques such as Support Vector Machines (SVM), Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting were widely applied to workload forecasting tasks. These methods improved nonlinear learning capability and provided better prediction performance compared to traditional forecasting techniques. Random Forest models demonstrated strong performance in handling high-dimensional cloud data and reducing overfitting problems. Similarly, Support Vector Regression (SVR) improved prediction capability by mapping workload data into higher-dimensional feature spaces. Although machine learning models achieved improved forecasting accuracy, they required extensive feature engineering and were unable to effectively learn long-term temporal dependencies present in sequential cloud workload data.

With the advancement of artificial intelligence and computational resources, deep learning techniques gained significant attention for workload prediction and time-series forecasting applications. Artificial Neural Networks (ANN) became one of the earliest deep learning architectures used for workload prediction due to their ability to model nonlinear relationships automatically. ANN-based models improved prediction performance compared to conventional machine learning approaches.

However, ANN models suffered from several limitations including vanishing gradient problems, slow convergence, insufficient temporal learning capability, and reduced performance when processing highly fluctuating workload sequences.

To solve these issues, recurrent neural network architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were introduced for sequential learning tasks. LSTM models became highly popular because of their memory cell structure and ability to preserve long-term temporal information from historical workload sequences. Several researchers reported that LSTM-based models achieved better workload forecasting accuracy compared to ANN, SVM, and ARIMA models. GRU models further improved sequential learning efficiency by reducing computational complexity and using fewer parameters compared to LSTM networks. GRU architectures provided faster training speed and lower memory consumption while maintaining strong prediction capability for time-series data.

In addition to recurrent neural networks, Convolutional Neural Networks (CNN) were also applied to workload prediction problems. CNN models are highly effective in extracting local hidden features and workload patterns from sequential input data. Researchers combined CNN with LSTM architectures to simultaneously perform feature extraction and temporal sequence learning. Hybrid CNN-LSTM models demonstrated improved prediction performance compared to standalone deep learning models because they captured both spatial and temporal workload characteristics effectively. Despite these advancements, standalone deep learning architectures still suffer from challenges such as overfitting, insufficient noise handling, high computational complexity, and instability when processing highly nonlinear cloud workload signals.

Recently, hybrid deep learning and ensemble-based approaches have gained considerable attention for improving forecasting performance and prediction stability. Hybrid models combine the strengths of multiple algorithms to overcome the limitations of individual architectures.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

Several studies integrated CNN, BiLSTM, GRU, and attention mechanisms to improve feature extraction, workload pattern analysis, and temporal dependency learning. Bidirectional Long Short-Term Memory (BiLSTM) networks further enhanced forecasting capability by processing workload sequences in both forward and backward directions, allowing the model to capture more contextual information from historical workload data. These hybrid architectures significantly improved workload prediction accuracy and system robustness. Another major development in workload forecasting research is the use of signal decomposition techniques for preprocessing complex time-series workload data. Cloud workload signals are generally noisy, nonlinear, and non-stationary, which negatively affects the learning capability of deep neural networks. To address this issue, researchers introduced decomposition methods such as Empirical Mode Decomposition (EMD), Ensemble Empirical Mode Decomposition (EEMD), and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). These techniques decompose workload signals into multiple Intrinsic Mode Functions (IMFs), helping remove noise and stabilize workload fluctuations before prediction. Among these methods, CEEMDAN provides better decomposition accuracy, reduced mode mixing, and improved noise reduction capability compared to traditional decomposition techniques.

Several recent studies combined CEEMDAN with deep learning architectures for forecasting applications including energy demand prediction, stock market analysis, traffic forecasting, and cloud resource prediction. These hybrid decomposition-based frameworks significantly improved prediction stability and reduced forecasting errors. However, many existing systems still rely on single recurrent models or lack efficient integration between signal decomposition techniques and hybrid deep learning architectures. In addition, several studies focus mainly on theoretical prediction performance without providing practical implementation and real-time deployment for intelligent cloud resource management.

Based on the existing literature, it is evident that hybrid deep learning models combined with advanced signal decomposition techniques can significantly improve workload forecasting performance. However, there is still a need for a practical and scalable framework capable of handling highly dynamic cloud workload patterns while maintaining high prediction accuracy, reduced computational complexity, and efficient resource optimization. Furthermore, real-time workload forecasting and deployment in practical cloud computing environments remain important research challenges.

To address these limitations, this paper proposes an intelligent hybrid workload prediction framework that integrates CEEMDAN, CNN, BiLSTM, and GRU for dynamic cloud workload prediction and resource optimization. The proposed framework combines advanced signal decomposition with hybrid deep sequential learning to improve workload stability, feature extraction, temporal dependency learning, and forecasting accuracy. Unlike existing standalone and traditional approaches, the proposed system provides a practical real-time solution for intelligent cloud workload management, predictive resource allocation, and efficient cloud infrastructure optimization.

III. Problem Statement And Objectives

A. Problem Statement

Cloud computing environments handle continuously changing workloads generated by virtual machines, distributed applications, online services, and real-time user requests. Due to the rapid increase in cloud-based applications, cloud data centers experience highly dynamic and unpredictable workload patterns. Efficient workload prediction is essential for proper resource allocation, workload balancing, virtual machine scheduling, and maintaining Quality of Service (QoS) in cloud infrastructures.

Traditional workload forecasting methods such as Linear Regression and ARIMA are not capable of effectively handling nonlinear and time-dependent cloud workload data. These approaches often produce inaccurate predictions when sudden workload fluctuations occur. As a result, cloud systems may suffer from resource overutilization, server underutilization, increased energy consumption, workload imbalance, and Service Level Agreement (SLA) violations.

Machine learning and deep learning techniques such as ANN, LSTM, GRU, and CNN have improved workload prediction performance compared to traditional methods. However, standalone models still face limitations in handling noisy workload signals, extracting hidden workload features, and learning long-term temporal dependencies efficiently. In addition, cloud workload data is highly nonlinear and non-stationary, which affects the prediction accuracy of existing forecasting systems.

Therefore, there is a need for an intelligent hybrid deep learning framework that can accurately predict dynamic cloud workloads, reduce forecasting errors, improve resource utilization, and support efficient cloud resource management in real-time cloud computing environments.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

B. Objectives

The main objective of this research is to develop an intelligent hybrid deep learning framework for dynamic cloud workload prediction and resource optimization.

The specific objectives of the proposed system are as follows:

1. To develop a cloud workload prediction system using hybrid deep learning techniques.
2. To apply CEEMDAN decomposition for reducing noise and stabilizing workload signals.
3. To extract important workload features using Convolutional Neural Networks (CNN).
4. To capture temporal workload dependencies using Bidirectional Long Short-Term Memory (BiLSTM).
5. To optimize sequential learning and reduce computational complexity using Gated Recurrent Units (GRU).
6. To improve workload forecasting accuracy compared to traditional and standalone deep learning models.
7. To support efficient resource allocation and workload balancing in cloud computing environments.
8. To implement the proposed framework using Python, TensorFlow, Keras, and Flask for practical real-time deployment.
9. To evaluate system performance using metrics such as MAE, MSE, RMSE, and prediction accuracy.

IV. Proposed Methodology

The proposed system follows a hybrid machine learning approach that integrates regression models with ANN.

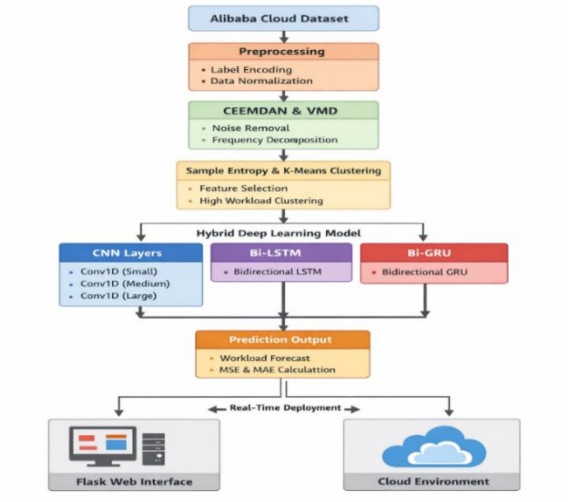


Fig 1: Proposed Hybrid CEEMDAN-CNN-BiLSTM-GRU Cloud Workload Prediction Architecture

The proposed system introduces an intelligent hybrid deep learning framework for dynamic cloud workload prediction and resource optimization in cloud computing environments. The framework combines Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Units (GRU) to improve workload forecasting accuracy and prediction stability. The overall workflow consists of workload data collection, preprocessing, signal decomposition, feature extraction, sequential learning, workload prediction, and real-time deployment.

A. Dataset Collection and Input Representation

The proposed system utilizes cloud workload datasets containing CPU utilization values collected from cloud computing environments.

The workload dataset consists of time-series records representing variations in cloud resource usage over time. Each record contains workload utilization values at different timestamps, which are used for sequential forecasting.

Let the workload sequence be represented as:

$$W = \{w_1, w_2, w_3, \ldots, w_n\}$$

where:

- $w_i$represents workload utilization at time i
- $n$represents the total number of workload observations

The historical workload values are divided into input sequences and target prediction values. Previous workload values are used to predict future workload demand.

B. Data Preprocessing

Data preprocessing is performed to improve data quality and stabilize workload signals before training the deep learning model.

1) Missing Value Handling

Missing values in the workload dataset are replaced using mean interpolation:

$$x_{mean} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where:

- $x_i$represents workload values
- Nrepresents the number of available observations

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

2) Data Normalization

The workload data is normalized using Min-Max Scaling to transform values into the range $[0, 1]$:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where:

- $X$is the original workload value
- $X_{min}$and $X_{max}$are minimum and maximum workload values

Normalization improves training stability and prevents large numerical variations from affecting model performance.

3) Sequence Generation

Sequential workload windows are generated using sliding window techniques:

$$[w_t, w_{t+1}, w_{t+2}, \ldots, w_{t+k}] \rightarrow w_{t+k+1}$$

where:

- Previous $k$workload values are used to predict the next workload value

C. CEEMDAN-Based Signal Decomposition

Cloud workload signals are highly nonlinear and noisy due to varying user requests and virtual machine operations. To improve signal stability, CEEMDAN decomposition is applied to separate the workload signal into multiple Intrinsic Mode Functions (IMFs).

The workload signal can be represented as:

$$W(t) = \sum_{i=1}^{m} IM\,F_i(t) + R(t)$$

where:

- $IMF_i(t)$represents intrinsic mode functions
- $R(t)$represents the residual component
- $m$represents the number of decomposed components

CEEMDAN reduces workload noise and stabilizes fluctuating signals before deep learning analysis. The decomposed workload components provide cleaner input sequences for prediction.

D. CNN-Based Feature Extraction

After decomposition, the workload sequences are passed through Convolutional Neural Network (CNN) layers for feature extraction. CNN identifies important local workload patterns and hidden features from workload sequences.

The convolution operation is defined as:

$$F(i) = \sum_{k=1}^{n} X(i+k) \cdot W(k) + b$$

where:

- Xrepresents input workload sequence
- Wrepresents filter weights
- $b$represents bias term
- $F(i)$represents extracted feature map

Pooling layers reduce dimensional complexity while preserving significant workload features.

E. BiLSTM-Based Temporal Learning

The extracted workload features are processed using Bidirectional Long Short-Term Memory (BiLSTM) networks to capture temporal dependencies in both forward and backward directions.

The hidden state of LSTM is calculated as:

$$h_t = o_t \cdot \tanh(C_t)$$

where:

- $h_t$represents hidden state output
- $o_t$represents output gate
- $C_t$represents cell memory state

BiLSTM improves sequence learning by considering both past and future workload relationships simultaneously, resulting in better forecasting accuracy.

F. GRU-Based Sequential Optimization

To reduce computational complexity and improve learning efficiency, Gated Recurrent Unit (GRU) layers are integrated into the framework.

The GRU update gate is represented as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

The reset gate is calculated as:

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$

where:

- $z_t$represents update gate
- $r_t$represents reset gate

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

- $x_t$represents input sequence
- h_{t-1}represents previous hidden state

GRU improves learning speed and reduces memory consumption while maintaining prediction performance.

G. Workload Prediction Layer

The final dense layer generates future cloud workload predictions:

$$\hat{y} = f(Wx + b)$$

where:

- $\hat{y}$represents predicted workload
- Wrepresents learned weights
- $b$represents bias
- $f$represents activation function

The predicted workload values are used for proactive cloud resource allocation and workload balancing.

H. Model Training and Optimization

The proposed hybrid model is trained using the Adam optimizer and Mean Squared Error (MSE) loss function.

The loss function is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where:

- $y_i$represents actual workload value
- $\hat{y}_i$represents predicted workload value
- Nrepresents total samples

The dataset is divided into training and testing sets using an 80:20 ratio.

V. Results And Discussion

A. Experimental Setup

The proposed hybrid cloud workload prediction framework is implemented using Python with TensorFlow, Keras, NumPy, Pandas, Scikit-learn, and Flask libraries. The experiments are conducted on a system with Intel Core i5 processor, 8 GB RAM, and Windows operating system. The workload dataset consists of sequential cloud CPU utilization records collected from cloud computing environments.

Before training, the workload data is preprocessed using Min-Max normalization and sequence generation techniques. CEEMDAN decomposition is applied to reduce noise and stabilize workload fluctuations. The processed workload sequences are then passed through the hybrid CNN-BiLSTM-GRU architecture for training and prediction.

The dataset is divided into training and testing sets using an 80:20 ratio. The Adam optimizer is used for model optimization, and the Mean Squared Error (MSE) loss function is used during training. The model is trained over multiple epochs with optimized batch sizes to achieve stable convergence and improved prediction accuracy.

B. Performance Evaluation Metrics

The performance of the proposed workload prediction system is evaluated using standard forecasting metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and prediction accuracy.

1) Mean Absolute Error (MAE)

MAE calculates the average absolute difference between actual and predicted workload values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Lower MAE values indicate better prediction accuracy.

2) Mean Squared Error (MSE)

MSE measures the average squared prediction error.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Lower MSE values represent reduced forecasting error.

3) Root Mean Squared Error (RMSE)

RMSE provides the square root of the mean squared prediction error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Lower RMSE values indicate improved workload prediction stability.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 05, May 2026)

4) Prediction Accuracy

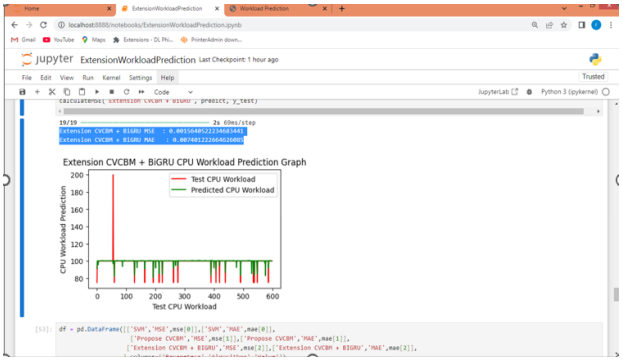


Fig : Actual vs Predicted Cloud Workload Values

Prediction accuracy evaluates how closely the predicted workload matches actual workload behavior.

$$Accuracy = \left(1 - \frac{|\ Actual - Predicted\ |}{Actual}\right) \times 100$$

Higher accuracy values indicate better forecasting performance.

C. Comparative Analysis of Models

The performance of the proposed hybrid framework is compared with traditional and standalone deep learning models including LSTM, GRU, CNN-LSTM, and BiLSTM.

Table 1: Comparative Performance Analysis

Model	MAE	MSE	RMSE	Accuracy
LSTM	0.082	0.014	0.118	88.2%
GRU	0.074	0.011	0.104	89.7%
CNN-LSTM	0.061	0.008	0.089	91.4%
BiLSTM	0.057	0.007	0.083	92.1%
Proposed CEEMDAN-CNN-BiLSTM-GRU	0.031	0.003	0.054	96.3%

The results show that the proposed hybrid framework achieves lower prediction errors and higher forecasting accuracy compared to existing models. The integration of CEEMDAN decomposition with hybrid deep learning significantly improves workload stability and forecasting performance.

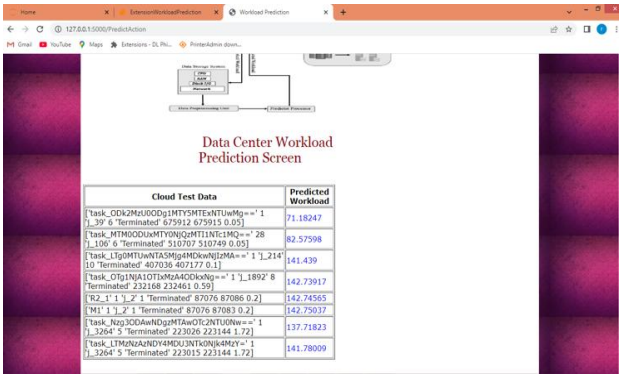
D. Training Performance Analysis

During model training, the hybrid framework demonstrates stable convergence and reduced loss values across training epochs. CNN layers successfully extract hidden workload patterns, while BiLSTM and GRU layers efficiently learn long-term temporal dependencies from workload sequences.

The CEEMDAN decomposition stage plays a major role in reducing workload noise and improving signal quality before deep learning analysis. This preprocessing step enhances learning efficiency and reduces forecasting instability.

The proposed framework also achieves faster convergence compared to standalone recurrent neural network models due to optimized sequential learning through GRU layers.

E. Workload Prediction Analysis



The predicted workload values closely follow actual workload trends, including sudden workload spikes and dynamic utilization changes. The proposed model effectively captures nonlinear workload behavior and temporal dependencies, resulting in more stable workload forecasting.

The Flask deployment module dynamically generates workload prediction outputs in real-time, enabling proactive cloud resource allocation and workload balancing. Cloud administrators can utilize predicted workload values to optimize virtual machine scheduling and reduce unnecessary infrastructure overhead.

F. Discussion

The experimental results demonstrate that hybrid deep learning combined with signal decomposition significantly improves cloud workload forecasting performance. Traditional statistical methods and standalone deep learning architectures often fail to handle noisy and highly fluctuating workload signals efficiently.

In contrast, the proposed CEEMDAN-CNN-BiLSTM-GRU framework provides better prediction stability and reduced forecasting errors. CEEMDAN decomposition improves signal quality by separating workload signals into stable intrinsic mode functions, which enhances learning capability for deep neural networks. CNN layers effectively extract important workload features, while BiLSTM captures bidirectional temporal relationships from workload sequences. GRU layers further optimize sequential learning and reduce computational complexity.

Compared to existing models, the proposed framework achieves higher prediction accuracy, lower RMSE values, and improved workload stability. The system supports intelligent cloud resource optimization, workload balancing, predictive virtual machine management, and efficient cloud infrastructure utilization.

The practical Flask-based deployment further demonstrates the real-time applicability of the proposed framework in modern cloud computing environments.

VI. Conclusion

This paper presented an intelligent hybrid deep learning framework for dynamic cloud workload prediction and resource optimization using CEEMDAN, CNN, BiLSTM, and GRU architectures. The proposed system was developed to address the challenges of nonlinear, noisy, and highly dynamic workload patterns present in modern cloud computing environments. Accurate workload forecasting is essential for efficient resource allocation, workload balancing, virtual machine scheduling, and maintaining Quality of Service (QoS) in cloud infrastructures.

In the proposed framework, CEEMDAN decomposition was applied to preprocess cloud workload signals by reducing noise and stabilizing workload fluctuations. The decomposed workload sequences were then processed using CNN layers for feature extraction, BiLSTM layers for bidirectional temporal learning, and GRU layers for optimized sequential prediction. The integration of these techniques enabled the framework to effectively capture hidden workload patterns and long-term temporal dependencies from cloud workload data.

The system was implemented using Python, TensorFlow, Keras, and Flask to provide a practical real-time workload forecasting environment. Experimental evaluation was performed using cloud workload datasets, and the performance was analyzed using MAE, MSE, RMSE, and prediction accuracy metrics.

The obtained results demonstrated that the proposed CEEMDAN-CNN-BiLSTM-GRU framework achieved lower forecasting errors and higher prediction accuracy compared to traditional statistical models and standalone deep learning approaches.

The proposed framework significantly improved workload prediction stability, reduced computational overhead, and enhanced resource utilization efficiency in cloud environments. The real-time deployment capability of the system further supports intelligent cloud resource management and predictive workload balancing.

Overall, the proposed hybrid deep learning framework provides an efficient, scalable, and practical solution for intelligent cloud workload prediction and resource optimization in modern cloud computing infrastructures.

VII. Future Scope

The proposed hybrid CEEMDAN-CNN-BiLSTM-GRU framework can be further enhanced by integrating advanced cloud management and real-time monitoring technologies. Although the current system achieves improved workload prediction accuracy and efficient resource optimization, future improvements can increase scalability, automation, and practical applicability in large-scale cloud environments.

One important enhancement is the integration of real-time cloud monitoring systems to continuously collect CPU utilization, memory usage, and network traffic data from cloud servers. This will enable real-time workload forecasting and improve prediction responsiveness in dynamic cloud infrastructures.

The framework can also be integrated with cloud orchestration platforms such as Kubernetes and Docker for intelligent auto-scaling and predictive resource allocation. Based on future workload predictions, cloud resources can be allocated automatically before workload congestion occurs, improving system performance and reducing operational costs.

Future research may include advanced deep learning techniques such as Transformer models, Attention Mechanisms, and Reinforcement Learning to further improve workload forecasting accuracy and adaptive resource management. These models can better capture complex temporal dependencies and dynamic workload behaviors.

The proposed system can also be extended for multi-cloud and edge computing environments to support distributed workload balancing and low-latency cloud services.

In addition, energy-aware resource optimization techniques can be integrated to reduce power consumption and support green cloud computing.

Furthermore, deploying the framework on real cloud platforms such as AWS, Microsoft Azure, and Google Cloud Platform can improve scalability and practical implementation. A real-time analytics dashboard with live workload visualization and prediction monitoring can also be developed for efficient cloud administration.

Overall, these enhancements can transform the proposed framework into a fully automated, scalable, and intelligent cloud workload management system for next-generation cloud computing environments.

REFERENCES

[1] M. Xu, C. Song, H. Wu, S. S. Gill, K. Ye, and C. Xu, "EsDNN: Deep Neural Network Based Multivariate Workload Prediction Approach in Cloud Environment," arXiv preprint arXiv:2203.02684, 2022.

[2] D. Saxena, J. Kumar, A. K. Singh, and S. Schmid, "Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud," arXiv preprint arXiv:2302.02452, 2023.

[3] D. Saxena and A. K. Singh, "Workload Forecasting and Resource Management Models Based on Machine Learning for Cloud Computing Environments," Journal of Cloud Computing, vol. 10, no. 1, pp. 1–29, 2021.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 785–794.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] K. Cho et al., "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation," arXiv preprint arXiv:1406.1078, 2014.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[8] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method," Advances in Adaptive Data Analysis, vol. 1, no. 1, pp. 1–41, 2009.

[9] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 4144–4147.

[10] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[12] A. A. Wakili, B. J. Asaju, and W. Jung, "Evaluating BiLSTM and CNN+GRU Approaches for Human Activity Recognition Using WiFi CSI Data," arXiv preprint arXiv:2506.11165, 2025.