Received: 26 January 2023 | Revised: 8 August 2023 | Accepted: 11 September 2023

DOI: 10.1002/dac.5634

RESEARCH ARTICLE

WILEY

Workload prediction for enhancing power efficiency of cloud data centers using optimized self-attention-based progressive generative adversarial network

G. Saravanan[1] | **A. V. Santhosh Babu**[2]

[1]Department of Artificial Intelligence and Data Science, Erode Sengunthar Engineering College, Perundurai, Erode, Tamilnadu, India

[2]Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Namakkal, Tamilnadu, India

Correspondence
G. Saravanan, Department of Artificial Intelligence and Data Science, Erode Sengunthar Engineering College, Perundurai, Erode,Tamilnadu, India.
Email: gsaravanan.pacet@gmail.com

Summary

Nowadays, future workload prediction is an important requirement in cloud data centers to maintain flexibility and scalability of resources. However, due to unexpected peaks, drops in workload, noise, and redundancy in user requests, there is a considerable variance in resource demands, making it difficult to accurately predict workloads. Therefore, a self-attention-based progressive generative adversarial network (SAPGAN) optimized with Giza Pyramids Construction Algorithm (GPCA)-based workload prediction is proposed for Sustainable Cloud Data Centers (CDC). At first, redundant data in historical data obtained through CDC are filtered utilizing Markov chain random field (MCRF) co-simulation method. These pre-processed historical data are supplied to the SAPGAN. The SAPGAN weight parameters are optimized by GPCA. The proposed method is analyzed using 2 benchmark datasets: HTTP traces from Saskatchewan and NASA. The simulation is implemented in JAVA. The performance metrics is examined to verify the efficacy of the proposed technique. The performance of the proposed approach provides 28.70%, 11.87%, and 14.79% higher accuracy; 30.15%, 11.72%, and 18.34% lesser energy consume for the dataset of NASA; and 5.32%, 2.45%, and 5.67% higher accuracy; 12.36%, 24.24%, and 34.16% lesser energy consume for the dataset of Saskatchewan HTTP traces compared with existing methods, such as auto-adaptive learning in a dynamic cloud environment (AADEA-WLP-CDC), a neural network model depending on biphase adaptive learning for anticipating cloud data center workload (BALNN-WLP-CDC) and multiple scale ensemble of deep learning framework for multistep-ahead cloud workload prediction (EMD-LSTM-GAN-WLP-CDC).

KEYWORDS

cloud computing, cloud data center, Giza Pyramids Construction Algorithm, Markov chain random field, workload prediction

1 | INTRODUCTION

Cloud computing (CC) is a widely used computing paradigms.[1] With service level agreements (SLAs) among cloud service providers (CSPs) and consumers, CC guarantees on-demand supply of computing, storage including network resources.[2] Workloads spike in response to simultaneous user demands; therefore, there may not be enough resources. Besides, when workloads remain low, they enter the idle state that leads to wasting resources.[3] Variations in workload result in over- or under-provisioning of resources, and it cause needless overheads or worst SLAs.[4] But, CSPs can compute the resource provisioning methods rapidly for ensuring SLAs by enhancing resource utilization.[5] The rapid and flexible workload prediction models are crucial for CC to achieve these goals.[6] By configuring and allocating resources in advance, a proficient and logical resource provisioning is achieved with respect to effective prediction for future workloads.[7] Workload prediction faces two challenges in CC; they are higher variation of workload patterns and higher dimensionality of workload data.[8]

To construct an exact workload prediction approach that handle extremely variable workloads, the workload pattern correlations must be successfully collected in response to highly variance of workload patterns.[9] The original workload data features are extracted to decrease the workload data dimensionality as well as prediction faults for precise workload prediction to overcome the high dimensionality difficulty.[10]

The challenge workload prediction has drawn a lot of scientific attention.[11] However, to accomplish accurate workload prediction, many traditional models are depending on regression theories, heuristics, or typical neural networks, which demand workloads with obvious regularity.[12] At the meantime, traditional neural networks do not fully utilize the correlation among neurons for improved prediction outcome.[13] So, they could not accurately estimate high variable workloads.[14] The majority of these models concentrate on workloads on small-scale grid or high performance computing schemes, which reveals less variance when analyzed to the huge-scale Sustainable Cloud Data Centers (CDC).[15] These models do not adjust to the real-world CC with high variable workloads, resulting severe dilapidation in the accuracy of workload prediction.[16]

To overwhelm these challenges in workload prediction, first historical data is received from CDC contains redundant data and noise are Preprocessed using Markov chain random field (MCRF) for predicting the workload self-attention-based progressive generative adversarial network (SAPGAN) is used. Next, Giza Pyramids Construction Algorithm (GPCA) is proposed to optimize the SAPGAN weight parameters.

The key contributions of this work are abridged below:

- MCRF co-simulation method is suggested for effectively pre-processing the higher dimensional input workload data that results in the actual workload compression by pre-processing the input units with higher level activation degrees.
- SAPGAN base prediction approach for cloud workloads is proposed to adapt high variable workloads and acquire exact workload prediction by captured the necessary historical data.
- The simulation experiments utilizing real-world workload traces are implemented to confirm the efficacy of proposed SAPGAN-GPCA-WLP-CDC-SHTD on the prediction of cloud workload. The outcomes prove that the SAPGAN-GPCA-WLP-CDC-SHTD outperforms the existing models such as AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD, respectively.

The remaining segments of this manuscript are organized as follows: Section 2 delineates the literature survey, Section 3 illustrates the proposed method, Section 4 proves the result and discussion, and finally, the conclusion is given in Section 5.

2 | LITERATURE SURVEY

Among the numerous studies on workload prediction in CDC, a few recent studies are revised in this part.

Saxena and Singh[17] introduced the forecasting workload using auto-adaptive learning in a dynamic cloud environment. This workload prediction approach uses an adaptive neural network to estimate typical workload over the series of prediction intervals. The adaptively learns workload traces for a specified forecast interval from previous data using Auto Adaptive Differential Evolution approach. NASA and Saskatchewan HTTP traces were the two benchmark datasets for assessing the performance of the suggested method. It provides higher accuracy with maximum energy consumption.

Kumar et al.[18] have suggested a neural network model depending upon biphase adaptive learning for anticipating cloud data center workload. The presented framework for workload forecasting used supervised learning and neural network. To increase the learning effectiveness of the predictive model, an updated and adaptive differential evolution algorithm was created. The program has the ability to optimize the most appropriate crossover and mutation operators. Owing to its adaptive nature from sampled data in pattern learning, it was seen that the learning's prediction accuracy and convergence rate have improved. It provides better accuracy with minimum correlation coefficient (CoC).

Yazdanian and Sharifian[19] have presented a multiscale ensemble of deep learning for multistep-ahead cloud workload prediction. The longer term nonlinear relationships of cloud workload time-series can be efficiently exploited in ensemble GAN/LSTM that utilizes stacked LSTM blocks. It was true for the high-frequency, noise-like components. It provides higher cloud workload prediction accurateness but more squared prediction error.

Al-Asaly et al.[20] have presented a method for resource allocation on autonomous CC that was based upon deep learning. The deep learning-based intelligent workload predicting method was presented for the provisioning of cloud resource. An effective deep learning depending upon diffusion convolution recurrent neural network was suggested to forecast future demand for CPU utilization, then define how to respond to the fluctuations of workload at the interval. It provides minimum energy consumption with lower recall.

Khan et al.[21] have presented a workload forecasting and energy state estimation in cloud data centers: ML-centric approach. ML-based method was presented to forecast load and energy to support resource management decisions. To modeling workload predictions, Linear Regression, Ridge Regression, ARD Regression, Elastic Net, deep learning approach, viz Gated Recurrent Unit was considered. The predictions were scaled by root mean square error. It provides maximum sum of elasticity index with minimum CoC.

Singh et al.[22] have introduced a quantum method toward the adaptive estimation of cloud workloads. The workload prediction depending on evolutionary quantum neural network was presented for cloud data center. By encoding workload information into qubits and disseminating via the network, it takes benefit of computational efficacy of quantum scaling to proactively assess workload or resource requirements with maximized accuracy. It provides higher precision with maximum energy consumption.

Saxena et al.[23] have presented Online VM Prediction-Based Multi-Objective Load Balancing Framework for Resource Management at Cloud Data Center, which predicts servers resource utilization and balance the load accordingly. By reducing the number of active servers, facilitating VM migrations, and improving resource use, it aids power conservation. To reduce the performance deterioration brought on by under or overloaded servers, an online resource prediction scheme was created and installed on every virtual machine. To lessen network traffic and data center power consumption, a multi-objective VM placement and migration algorithms were presented. It presents greater precision but lower recall.

Jeddi and Sharifian[24] have presented a hybrid wavelet decomposer and GMDH-ELM ensemble method for Network function virtualization workload forecasting in CC. To present NFV resources and meet SLA conditions, energy consume was decreased and used physical resources. An optimized resource allocation considered resource auto-scaling property for NFV services on account of network traffic was changing rapidly. To scale cloud resources, forecast the NFV workload. It provides minimum mean square prediction error conversely with maximum predictions in error range (PERs).

3 | PROPOSED METHODOLOGY

To reduce operational costs and increase cloud efficiency, workload prediction is done in cloud data. Accuracy is the primary factor of workload prediction; however, the existing models are not attaining the adequate performance. To address these problems, the SAPGAN with task load prediction model based on the GPCA is proposed in this work. The proposed SAPGAN-GPCA-WLP-CDC model's primary goal is to upgrade the efficiency of power in cloud data centers. The historical data received from CDC contain redundant data that are filtered utilizing MCRF co-simulation method. These pre-processed historical data are fed to SAPGAN for predicting workload at dynamic cloud environment; also it generate predicted workload as output. Generally, SAPGAN does not adopt any optimization methods to determine the optimal parameters and ensuring accurate prediction. That's why, GPCA is employed to optimize the SAPGAN weight parameters. Finally, the proposed SAPGAN-GPCA-WLP-CDC approach exactly predicts the upcoming workload as well as decreases excess power consume in CDC. Figure 1 depicts the proposed SAPGAN-GPCA-WLP-CDC model.

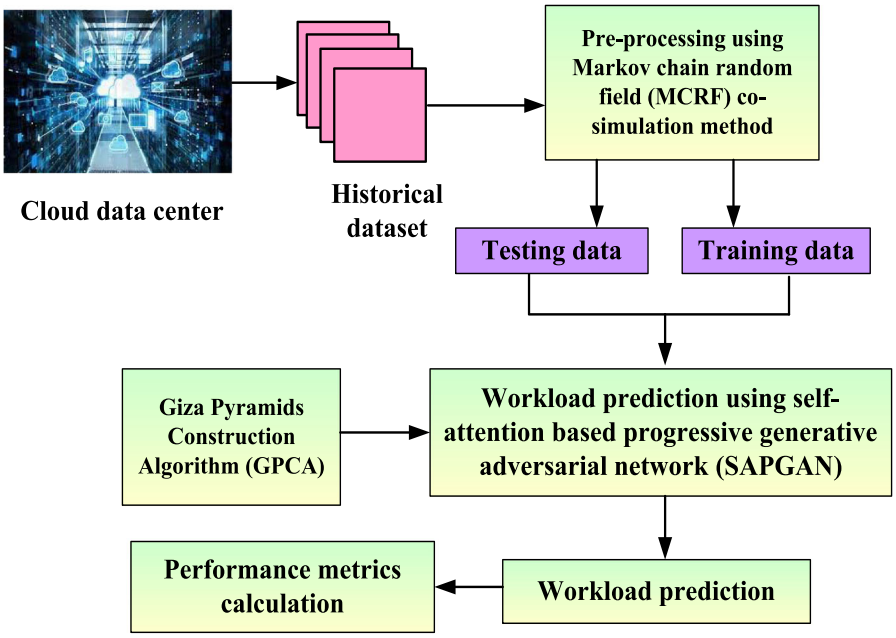

FIGURE 1 Proposed SAPGAN-GPCA-WLP-CDCmethodology.

3.1 | Data acquisition

NASA and Saskatchewan HTTP traces are the 2 benchmark data sets that are applied to validate the outcomes. The first log is gathered at 00:00:00 July 1, 1995 to 23:59:59 July 31, 1995, totally 31 days. NASA contains 1,891,715 HTTP requests. The Saskatchewan consists of a 7-month HTTP request log from a university WWW server. This log is gathered at 00:00:00 June 1, 1995 to 23:59:59 December 31, 1995, totally 214 days. At the period of 7 months, total request are 2,408,625. Note that the timestamp has a resolution of 1 s.

3.2 | Pre-processing using MCRF co-simulation method

This section describes the historical information from cloud data. This data received from CDC contain redundant data and noise that is filtered utilizing MCRF co-simulation method. Redundant data refer to information that is duplicated or repetitive within a dataset. It doesn't provide any additional value or insights and can be safely removed without affecting the analysis or results. Redundancy can occur due to various reasons, such as data collection errors, data merging processes, or technical issues, and cleaning noisy data is an important pre-processing step to improve the quality and integrity of the dataset. Pre-processing filters out noise and redundant data from the center using MCRF co-simulation method. Using a Markov chain, the result of combining the MCRF model with the histogram-oriented gradient is the random field characteristics that are employed to reduce noise as well as duplicated data on dataset. The cloud data center historical data contain unnecessary data. Unnecessary data refer to information that is not relevant or useful for the specific analysis or task at hand. It includes features or variables that do not contribute to the desired outcome or have negligible impact on the results. Removing unnecessary data can simplify the dataset, reduce computational overhead, and improve the efficiency of subsequent analyses. A Markov chain is a probabilistic model with a particular kind of dependence.

Let $J_0,J_1,J_2....J_m$ be the random variable of the sequence taking value in state space $\{R_1,R_2....R_n\}$. The sequence of a Markov process is described in Equation (1).

$$\begin{aligned} &Q_s\left(J_i=R_l|J_{i-1}=R_k,J_{i-2}=R_n,J_{i-3}=R_s...,J_0=R_q\right)= \\ &\quad Q_s(J_i=R_l|J_{i-1}=R_k)=q_{kl} \end{aligned} \tag{1}$$

where $\}|\}$ as symbol for conditional probability and $\}=\}$ is denoted as the notation convenience. Consider the Markov chain (A_i) and (B_i) is defined on the state space $\{R_1, R_2 .... R_n\}$, and the positive transition probability is defined as Equation (2).

$$Q_s\left(A_{i+1}=R_l, B_{j+1}=R_e | A_i=S_k, B_j=R_m\right)=q_{km,le} \tag{2}$$

The coupled transition probability $q_{km,le}$ on the state space $\{R_1, R_2 .... R_n\} \times \{R_1, R_2 .... R_n\}$ is given in Equation (3).

$$q_{km,le}=q_{kl} \cdot q_{me} \tag{3}$$

Two coupled one-dimensional Markov chain (A_i) and (B_i) is used to remove interference and is expressed in Equation (4),

$$\begin{aligned} Q_s\left(J_{i,j}=R_l \,|\, J_{i-1,j}=R_k, J_{i,j-1}=R_m\right)=D.Q_s(A_i=R_l|A_{i-1}=R_l) \cdot \\ Q_s\left(B_j=R_l|B_{j-1}=R_n\right) \end{aligned} \tag{4}$$

where D denotes the normalizing constant that arises by not permitting transitions in the (A_i) and (B_i) chain to different states, Q_s signifies function of random noise, and random noise refers to the presence of unpredictable and random fluctuations in data. Random noise can refer to irrelevant or unwanted variations that may occur due to various factors, such as measurement errors, environmental conditions, or other sources of interference. The extreme points with noise intrusion on input historic data are eradicated by Equation (5).

$$\widehat{y}_t=a_t-\theta_j \times a_{t-1} \tag{5}$$

Let $\widehat{y}_t$ as pre-processed historic data at time interval t, a_t as noise, θ_j implies weight parameters, and a_{t-1} specifies random noise. The noise along inappropriate data is eliminated from historical dataset utilizing MCRF co-simulation process. The filtering historic data are fed to neural network for workload estimation.

3.3 | SAPGAN-based workload prediction

A progressive SAPGAN technique is employed to forecast workload. The advantage of using SAPGAN for workload prediction in cloud data centers is that they capture temporal dependencies, handle variable-length sequences, generate realistic workload samples, and scale efficiently. The self-attention mechanism allows the model to capture complex temporal patterns, while the GAN framework generates realistic samples. The model can handle variable-length sequences without pre-processing and parallelizes well for scalability. Also, it allows the model to process the workload data in a parallelized manner, reducing the computational time required for prediction. Additionally, the learned representations are transferable to new data centers or domains, enabling accurate predictions in unseen scenarios. Overall, SAPGANs are a promising approach for efficient and accurate workload estimation in cloud data centers. By utilizing a neural network model, forecast the useful data from the samples of training input data. However, the data center for the cloud's workload changes frequently. The SAPGAN generates scheduled workload output for workload prediction in heterogeneous cloud. The SAPGAN is a kind of GAN; this is applied in numerous tasks, such as graph embedding, machine translation, visual recognition, and generative modeling. To investigate the features hierarchy, the convolution method is stacked into a number of layers, also acts as a basic component for each advanced machine vision structures. These representations have been learned during the series of convolution operations. The primary usage of convolution layers is in data processing. All the experiments had one thing in general: no one employed convolution techniques to capture geometric shapes.

The non-local block is a global context model since it incorporates query-specific global context attributes at each query location. The time including space complexity of non-local block is quadratic to m_q positioning and is generated for every query location. The non-local block may be denoted in Equation (6).

$$J_x = b_x + K_J \sum_{y=1}^{m_q} \frac{s(b_x, b_y)}{H(b)} (K_v . b_y) \tag{6}$$

Consider index of query positions refers x, and y counts all feasible locations. The relationship between position x and y is represented as b_x, b_y, $H(b)$ represents normalization factor, K_J and K_v represents linear transform matrices. For simplification, $w_{xy} = \frac{s(b_x, b_y)}{H(b)}$ as normalized pair wise the relationship between position x and y in the non-local block is revealed in Equation (7),

$$J_x = b_x + \sum_{y=1}^{m_q} \frac{Exp(K_r b_y)}{\sum_{m=1}^{m_q} Exp(K_r b_m)} (K_r b_y) \tag{7}$$

where K_r and K_v signify linear transformation matrix. They lessen computation cost in Equation (7) using distributive law to move K_v outer of attention pooling, expressed in Equation (8).

$$J_x = b_x + K_v \sum_{y=1}^{m_q} \frac{Exp(K_r b_y)}{\sum_{m=1}^{m_q} Exp(K_r b_n)} (b_y) \tag{8}$$

Equation (8) presents self-attention to the structure of progressive generative adversarial networks. The self-attention mechanism's inclusion instructs the progressive generative adversarial networks to emphasis on target constructions of diverse shapes with sizes. Self-care is consolidated previously the down sample layer of the discriminator, then the up sample layer of the generator. In this manuscript, GPCA is utilized to optimize SAPGAN weight parameters. Here, GPCA is utilized to tune weight and bias K_r and K_v parameters of SAPGAN.

3.4 | Optimize the SAPGAN parameters using GPCA

GPCA is proposed to enhance the SAPGAN weight parameters. GPCA is based on the antique inspired meta-heuristic algorithm. The GPCA is similarly as Giza Necropolis; it is a place that consists of big three pyramids; all are constructed through the fourth dynasty of antique Egypt. GPCA is derived by the activities of the labors and push the stone slabs on the slope. Hence, the location of some labors will be substitute with others. This substitute changes the stone slabs movement and the power balance. In the construction procedure, some labors are substitute and placed into a new location. The updating location of labors in GPCA is used to optimize weight and bias K_r and K_v parameters of SAPGAN. Figure 2 represents the flow chart of GPCA. The step-by-step process for GPCA is delineated beneath.

Step 1. Initialization

The population distribution of Giza Pyramids Construction is initialized uniformly in the solution space utilizing Equation (9).

$$Y_j, g_k(t^{\}}) = \left(y_j^c(t^{\}}), \ldots y_j^m(t^{\}}) + \eta_k nh \cos\theta\right) \tag{9}$$

Here, $Y_{j,k}(t^{\}})$ epitomizes initial population of GPCA with j^{th} position and body solution in $t^{\}}$ time, and g_k epitomizes kinetic resistance force for GPCA. Then n represents the stone slap masses, η_k represents the kinetic resistance coefficient, h represents the earth gravity, and θ represents the ramp angle of GPCA.

Step 2. Random generation

Afterward initialization process, generate the input parameters of GPCA randomly. The maximal fitness values are observed, and the selection of optimum path is dependent on fitness function.

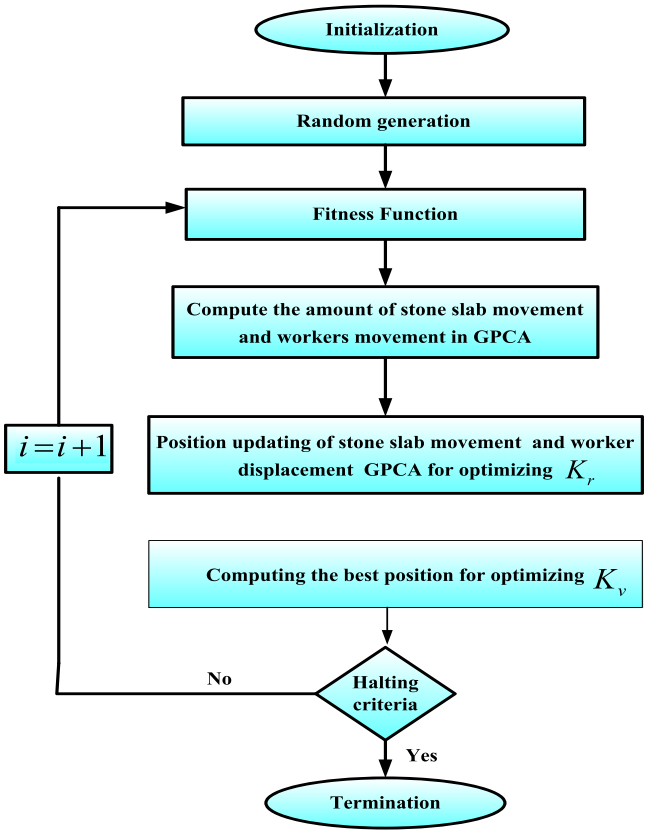

FIGURE 2 Flow chart of GPCA algorithm.

Step 3. Compute the fitness function

Create the random solution count through the initialized values. This is associated to weight with bias parameter K_r and K_v. Fitness function is calculated in Equation (10).

$$Fitness_{function} = optimization\ (K_r and K_v) \tag{10}$$

Step 4. Compute the amount of stone slab movement and workers movement in GPCA

The main point of GPCA is labors push the stone slab continuously moved to increase the feasible control and feasible lead of stone slab. This shock wave causes the employee to achieve non-repetitive displacement to pushing the stone slab well. Therefore, the stone movement on the slope can be calculated in Equation (11),

$$Displacement\ of\ stone\ slap = \frac{u_o^2}{2h(\sin\theta + \eta_k \cos\theta)} \tag{11}$$

where h represents the earth gravity and u_o represents the initial speed of the stone slab. The above expression is used to control the new position of the labors. Thus, the location of the labors pushes the stone slab is given in below expression (12),

$$Movement\ of\ workers = \frac{u_o^2}{2h\sin\theta} \tag{12}$$

where θ represents the ramp angle.

Step 5. Position updating of stone slab movement (trust node with low power level) and worker displacement (best node selection) GPCA for optimizing K_r

After computing the deviations of stone slab movement (trust path with low power level) and worker displacement (trust path selection) through the above Equations (11) and (12), a new location can be attained from the subsequent of the above expressions. The new location can be attained by adding the stone slab movement of exiting location that can be multiplied by the workers' movement. This updating location is a new resolution. Therefore, the updating position of stone slab and the workers can be expressed in below Equation (13).

$$\vec{Q} = \left(\vec{Q}_i + c\right) \times y\vec{\sigma}_i \tag{13}$$

where $\vec{Q}_i$ represents the current location, c represents the movement of the stone slab, y indicates the displacement of labor, and $\vec{\sigma}_i$ represents the random vector which following the Uniform and Normal distribution. This worker movement is applied to assess the present better solution to reach ideal global solution that is the trust node is attained as well as eliminated the workload. This stone slab movement is used to predict the workload. From Equation (12), the updated position will maximize the K_r.

Step 6. Computing the best position for optimizing K_v

Every solution is stationary during each repetition, and separate exterior workers are displayed in the space referred to as exterior workers. This worker strikes with additional position and alters its location toward a best one. So compute the best position of workers for exterior workers lessened at time involving the extreme mass of union. The best positions of exterior workers are determined by Equations (14) and (15).

$$N\left(t^{\}}\right) = 1 - \frac{t^{\} } - 1}{T^{\} } - 1} \tag{14}$$

$$V^{(c)}{}_j\left(t^{\}}\right) = s_1 . \left(1 - \frac{t^{\} } - 1}{T^{\} } - 1}\right) . V_{\max} . sign^{\} }\left(y^c_{best}\left(t^{\}}\right) - y^c_j\left(t^{\}}\right)\right) \tag{15}$$

Here, $N\left(t^{\}}\right)$ denotes mass of solution along time $t^{\}}$, $T^{\}}$ denotes maximal count of iterations, $V^{(c)}{}_j\left(t^{\}}\right)$ denotes exterior workers speed including c dimension and j^{th} system body iteration, s_1 implicates random term, $V_{\max}$ implies reality of the random count, and $y^c_{best}\left(t^{\}}\right) and y^c_j\left(t^{\}}\right)$ implies system bodies c dimension including best fitness for j^{th} system body iteration.

Step 7. Termination

Stop the procedure after getting better solution in Equations (16)–(21) repeats until fulfill the conditions. The output of GPCA algorithm gives an ideal workload prediction, which iteratively repeat step 3 until fulfill the halting criteria $d = d + 1$.

4 | RESULTS AND DISCUSSION

In this section, the simulation of SAPGAN optimized with Giza Pyramids Construction approach-based workload prediction for maximizing CDC power efficiency (SAPGAN-GPCA-WLP-CDC) is discussed. The performance is analyzed under NASA including Saskatchewan HTTP traces. The experiment is done in JAVA on PC along Intel Core i5, 2.50-GHz central processing unit, 8GB RAM, Windows 7. The metrics is analyzed. The performance of the proposed method is compared with existing AADEA-WLP-CDC,[17] BALNN-WLP-CDC,[18] and EMD-LSTM-GAN-WLP-CDC[19] methods. The simulation setup is given in Table 1.

TABLE 1 Simulation setup.

Parameters	Values
Input neural node (p)	10
Hidden layer node (q)	7–20
Output layer node (r)	1
Maximal epoch $(G_{\max})$	250
Training data size	75%
Count of population (L)	15

4.1 | Dataset description

NASA as well as Saskatchewan HTTP traces are the 2 datasets that are applied to validate the outcomes. The ASCII files containing one HTTP request in a row each include the internet traffic traces. The busy WWW server at the NASA Kennedy Space Center in Florida, NASA traces have two months' worth of HTTP requests stored. The first log is gathered at 00:00:00 July 1, 1995 to 23:59:59 July 31, 1995, totally 31 days. NASA contains 1,891,715 HTTP requests. The Saskatchewan consists of a 7-month HTTP request log from a university WWW server. Host, Time stamp, HTTP request, HTTP reply, and Bytes sent in the reply are the attributes of the data set. This log is gathered at 00:00:00 June 1, 1995 to 23:59:59 December 31, 1995, totally 214 days. At the period of 7 months, there are 2,408,625 requests. Note that the timestamp has a resolution of 1 s.

4.2 | Performance metrics

The proposed SAPGAN-GPCA-WLP-CDC approach performance is evaluated under accuracy, precision, recall, RMSE, energy consumption, CoC, MSPE, and PER.

4.2.1 | Accuracy

This is determined to validate the proposed SAPGAN-GPCA-WLP-CDC method efficiency for forecasting workload in the cloud system, which is calculated using Equation (16),

$$Accuracy = \frac{T(P) + T(N)}{T(P) + F(P) + T(N) + F(N)} \tag{16}$$

where $T(P)$ implies true positive, $F(P)$ implicates false positive, $T(N)$ refers true negative, and $F(N)$ signifies false negative.

4.2.2 | Precision

It is computed to scale the proposed method effectiveness with the help of Equation (17).

$$\text{Pr}ecision = \frac{T(P)}{T(P) + F(P)} \tag{17}$$

4.2.3 | Recall

This is defined to maximize power consume when forecasting workload and is computed by Equation (18).

$$\text{Recall} = \frac{T(P)}{T(P) + F(N)} \tag{18}$$

4.2.4 | Energy consumption

In the context of increasing power efficiency of cloud data centers, the energy efficiency formula for workload prediction can be defined as given in Equation (19).

$$Energy_{consumption} = (power\ consumption\ by\ workload)/(amount\ of\ workload\ accomplished) \tag{19}$$

This formula measures the ratio of power consumed by the workload to the amount of workload accomplished. It quantifies the energy efficiency of CDC in terms of power resources to accomplish the workload.

4.2.5 | CoC

The statistical relation is the computation of movement level amid the real and forecasted workload. This is scaled by Equation (20),

$$Correlation_{Coefficient} = C_F(W_a(T) \cdot W_P(T)) \tag{20}$$

where C_F denotes the correlation function and W_a and W_p denote real along predicted workloads at time period T.

4.2.6 | Mean squared prediction error (MSPE)

It contains higher accuracy when MPE score is closer to zero value; it is determined by Equation (21).

$$MSPE = 1/m \sum_{T=1}^{m} (W_a(T) - W_P(T))^2 \tag{21}$$

4.3 | Performance analysis of proposed approach for NASA dataset

Figures 3–8 show the simulation result of proposed approach for the dataset of NASA. The efficiency of the proposed technique is evaluated to the existing AADEA-WLP-CDC-ND,[17] BALNN-WLP-CDC-ND,[18] and EMD-LSTM-GAN-WLP-CDC-ND[19] methods.

Figure 3 signifies accuracy analysis for NASA dataset. In this figure, the SAPGAN-GPCA-WLP-CDC-ND method provides 21.97%, 20.39%, and 42.62% higher accuracy for prediction interval 20 min; 31.94%, 20.25%, and 33.80% higher accuracy for prediction interval 40 min; 22.66%, 40.43%, and 29.20% higher accuracy for prediction interval 60 min; 12.05%, 43.08%, and 36.76% higher accuracy for prediction interval 80 min; and 11.76%, 33.80%, and 41.79% higher accuracy for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND.

Figure 4 represents the precision analysis for the dataset of NASA. Here, the SAPGAN-GPCA-WLP-CDC-ND method provides 24.29%, 40.32%, and 14.47% higher precision for prediction interval 20 min; 19.74%, 31.88%, and 28.17% higher precision for prediction interval 40 min; 28.99%, 41.27%, and 30.88% higher precision for prediction interval 60 min; 28.77%, 40.30%, and 28.77% higher precision for prediction interval 80 min; and 24.00%, 60.34%, and 32.85%

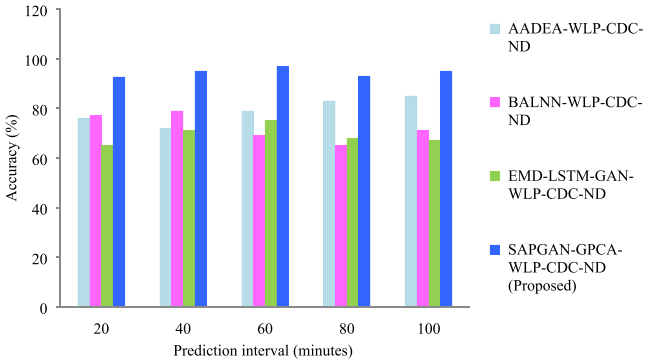

FIGURE 3 Performance analysis of accuracy for NASA dataset.

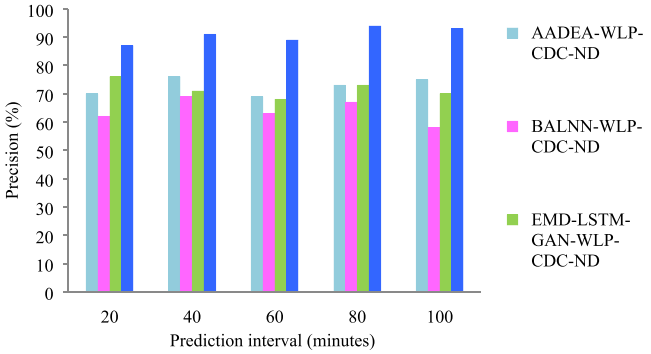

FIGURE 4 Performance analysis of precision for NASA dataset.

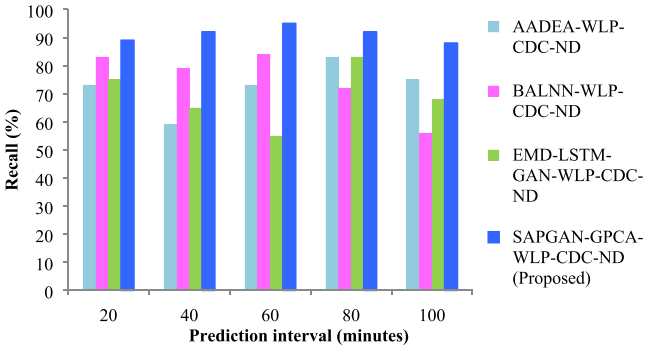

FIGURE 5 Performance analysis of recall for the dataset of NASA.

higher precision for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND, respectively.

Figure 5 implicates the recall analysis for NASA dataset. The proposed SAPGAN-GPCA-WLP-CDC-ND method provides 21.92%, 7.23%, and 18.67% higher recall for prediction interval 20 min; 55.93%, 16.46%, and 41.54% higher recall for prediction interval 40 min; 30.14%, 13.10%, and 72.73% higher recall for prediction interval 60 min; 13.70%, 15.28%, and 2.55% higher recall for prediction interval 80 min; and 17.33%, 57.14%, and 29.41% higher recall for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND, respectively.

Figure 6 represents the energy consumption analysis for NASA dataset. The proposed SAPGAN-GPCA-WLP-CDC-ND method provides 54.55%, 31.03%, and 49.38% lower energy consumption for prediction interval 20 min; 63.41%, 60.32%, and 70.59% lower energy consumption for prediction interval 40 min; 60.49%, 49.28%, and 59.77% lower energy consumption for prediction interval 60 min; 63.77%, 67.11%, and 68.42% lower energy consumption for prediction

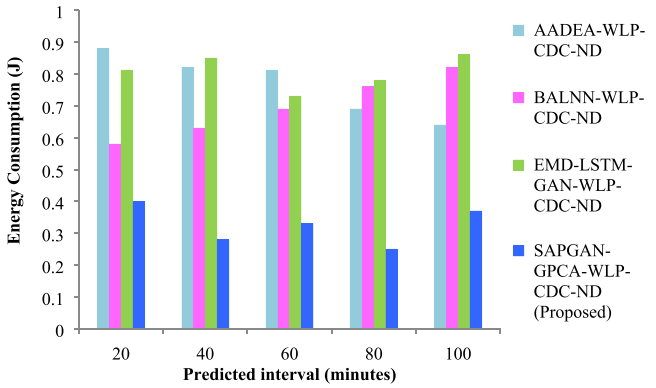

FIGURE 6 Performance analysis of energy consumption for NASA dataset.

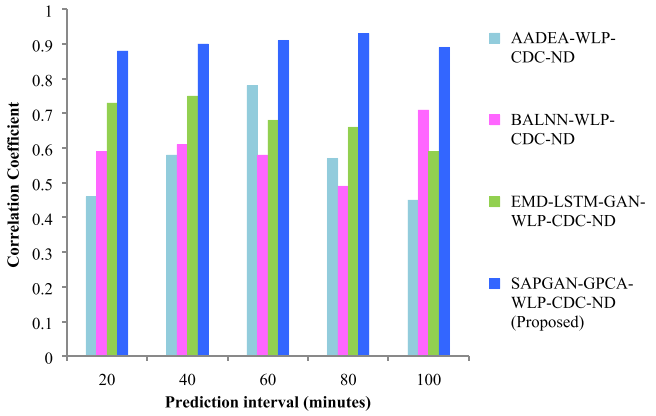

FIGURE 7 Performance analysis of correlation of coefficient for NASA dataset.

interval 80 min; and 42.19%, 58.54%, and 58.14% lower energy consumption for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND, respectively.

Figure 7 displays the correlation of coefficient analysis for NASA dataset. The proposed SAPGAN-GPCA-WLP-CDC-ND method provides 91.30%, 48.98%, and 20.55% higher correlation of coefficient for prediction interval 20 min; 55.17%, 47.54%, and 20.00% higher correlation of coefficient for prediction interval 40 min; 16.67%, 58.62%, and 33.82% higher correlation of coefficient for prediction interval 60 min; 62.28%, 89.80%, and 40.91% higher correlation of coefficient for prediction interval 80 min; and 97.78%, 25.35%, and 50.85% higher correlation of coefficient for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND, respectively.

Figure 8 shows the MSPEanalysis for NASA dataset. The proposed SAPGAN-GPCA-WLP-CDC-ND method provides 87.50%, 78.57%, and 40.00% lower MSPE for prediction interval 20 min; 78.95%, 76.92%, and 66.67% lower MSPE for prediction interval 40 min; 81.82%, 84.62%, and 88.89% lower MSPE for prediction interval 60 min; 75.00%, 80.00%, and 75.00% lower MSPE for prediction interval 80 min; 77.78%, 66.67%, and 54.55% lower MSPE for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-ND, BALNN-WLP-CDC-ND, and EMD-LSTM-GAN-WLP-CDC-ND, respectively.

4.4 | Performance analysis of proposed approach for Saskatchewan HTTP traces dataset

Figures 9–14 show the simulation result of proposed approach for the dataset of Saskatchewan HTTP traces. The proposed method performance is evaluated to the existing AADEA-WLP-CDC-SHTD,[17] BALNN-WLP-CDC-SHTD,[18] and EMD-LSTM-GAN-WLP-CDC-SHTD[19] methods.

Figure 9 displays the accuracy analysis for the dataset of Saskatchewan HTTP traces. Here, SAPGAN-GPCA-WLP-CDC-SHTD method provides 58.47%, 56.67%, and 27.94% higher accuracy for prediction interval 20 min; 22.78%, 29.85%, and 25.00% higher accuracy for prediction interval 40 min; 42.03%, 30.56%, and 63.33% higher accuracy for prediction interval 60 min; 54.55%, 39.34%, and 14.46% higher accuracy for prediction interval 80 min; and 40.00%, 31.43%, and 26.32% higher accuracy for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD, respectively.

Figure 10 represents the precision analysis for the dataset of Saskatchewan HTTP traces. Here, SAPGAN-GPCA-WLP-CDC-SHTD method provides 21.62%, 30.77%, and 20.00% higher precision for prediction interval 20 min; 20.51%, 27.78%, and 15.38% higher precision for prediction interval 40 min; 20.83%, 28.36%, and 19.18% higher precision for

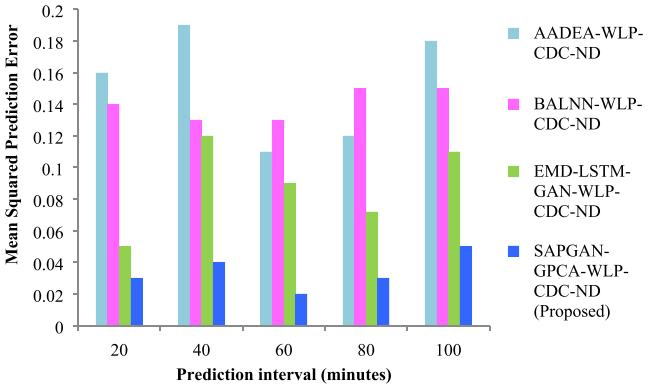

FIGURE 8 Performance analysis of MSPE for the dataset of NASA.

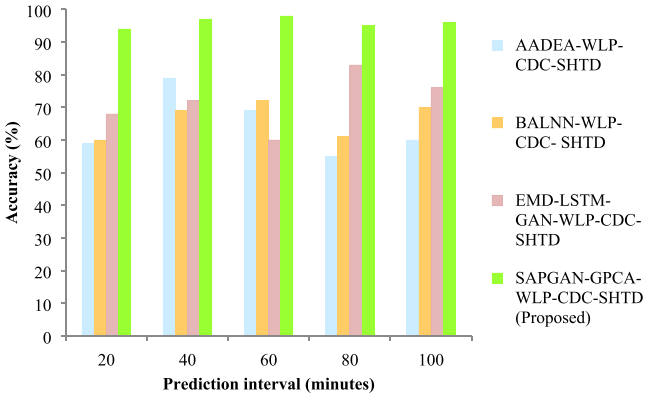

FIGURE 9 Performance analysis of accuracy for Saskatchewan HTTP traces dataset.

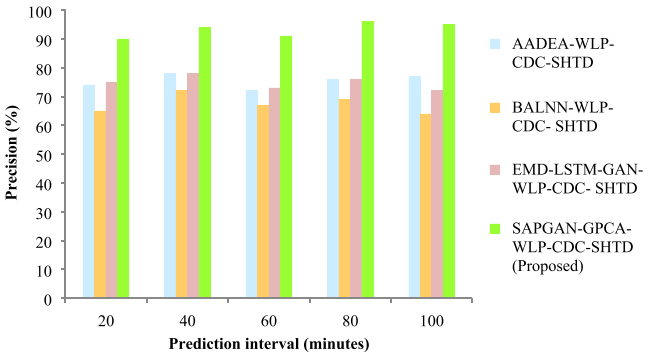

FIGURE 10 Performance analysis of precision for the dataset of Saskatchewan HTTP traces.

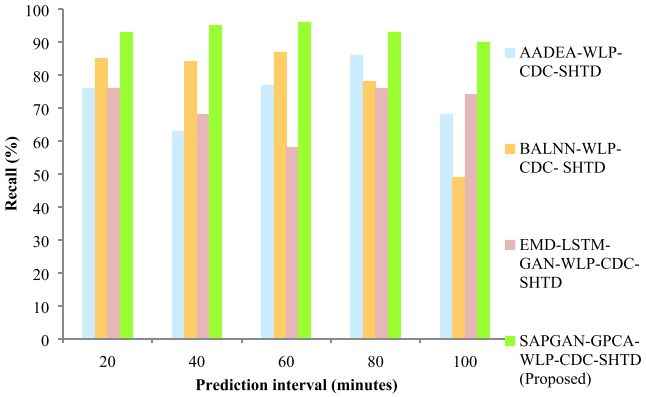

FIGURE 11 Performance analysis of recall for the dataset of Saskatchewan HTTP traces.

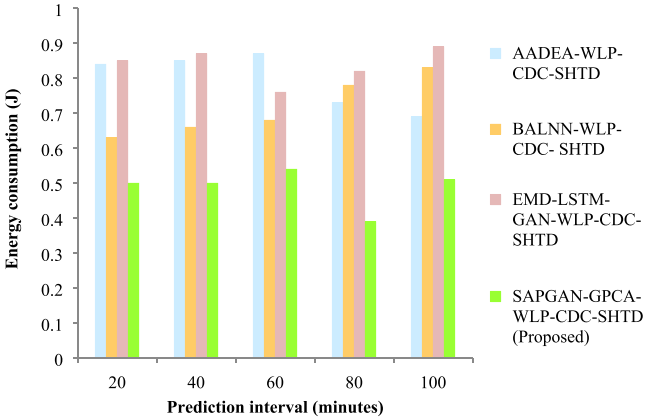

FIGURE 12 Performance analysis of energy consumption for the dataset of Saskatchewan HTTP traces.

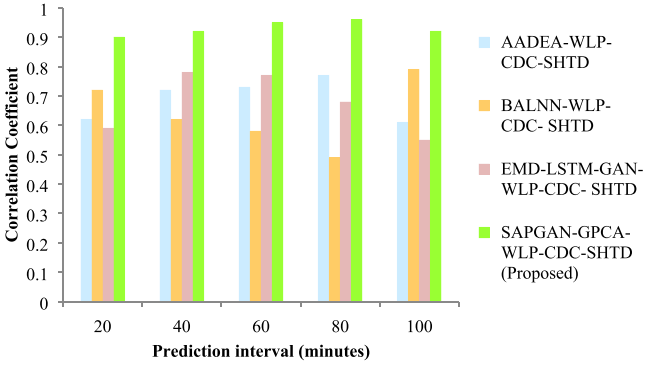

FIGURE 13 Performance analysis of correlation of coefficient for Saskatchewan HTTP traces data set.

prediction interval 60 min; 26.32%, 26.09%, and 26.32% higher precision for prediction interval 80 min; and 22.08%, 32.81%, and 25.00% higher precision for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD respectively.

Figure 11 portrays recall analysis for the dataset of Saskatchewan HTTP traces. The SAPGAN-GPCA-WLP-CDC-SHTD method provides 22.37%, 9.41%, and 21.05% higher recall for prediction interval 20 min; 33.33%, 11.90%, and 29.41% higher recall for prediction interval 40 min; 20.78%, 10.34%, and 37.93% higher recall for prediction interval 60 min; 8.14%, 15.38%, and 22.37% higher recall for prediction interval 80 min; 26.47%, 48.98%, and 20.27% higher recall

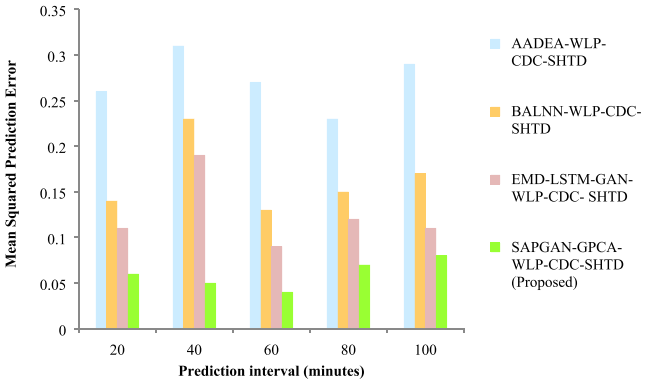

FIGURE 14 Performance analysis of MPE for the dataset of Saskatchewan HTTP traces.

TABLE 2 Comparative table for NASA dataset.

Author name	Accuracy (%)	Precision (%)	Recall (%)	Energy consumption (J)	Correlation coefficient	Mean squared prediction error
Saxena and Singh[17]	69.55	75.35	55.58	0.86	0.48	0.18
Kumar et al.[18]	68.33	70.43	68.44	0.64	0.53	0.14
Yazdanian and Sharifian[19]	79.92	66.85	59.08	0.82	0.54	0.16
Al-Asaly et al.[20]	63.12	-	60.32	0.55	0.86	0.13
Khan et al.[21]	60.32	72.13	84.22	-	0.44	0.08
Singh et al.[22]	72.53	70.23	-	0.87	0.76	-
Saxena et al.[23]	-	63.68	53.34	0.53	0.67	0.06
Jeddi and Sharifian[24]	88.64	67.11	88.82	-	0.86	0.1
SAPGAN-GPCA-WLP-CDC-ND (proposed)	99.48	99.84	99.74	0.42	0.9	0.02

for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD.

Figure 12 displays the energy consumption analysis for the dataset of Saskatchewan HTTP traces. The proposed SAPGAN-GPCA-WLP-CDC-SHTD method provides 50.00%, 20.63%, and 58.82% lower energy consumption for prediction interval 20 min; 58.82%, 24.24%, and 62.07% lower energy consumption for prediction interval 40 min; 39.08%, 20.59%, and 32.89% lower energy consumption for prediction interval 60 min; 54.79%, 50.00%, and 54.88% lower energy consumption for prediction interval 80 min; 26.09%, 38.55%, and 45.19% lower energy consumption for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD, respectively.

Figure 13 shows correlation of coefficient analysis. Here, SAPGAN-GPCA-WLP-CDC-SHTD method provides 45.16%, 25.00%, and 52.54% higher correlation of coefficient for prediction interval 20 min; 27.78%, 48.39%, and 17.95% higher correlation of coefficient for prediction interval 40 min; 30.14%, 63.79%, and 24.68% higher correlation of coefficient for prediction interval 60 min; 24.68%, 95.92%, and 41.18% higher correlation of coefficient for prediction interval 80 min; 50.82%, 16.46%, and 67.27% higher correlation of coefficient for prediction interval 100 min compared with existing methods like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-GAN-WLP-CDC-SHTD, respectively.

Figure 14 shows the performance analysis of MPE. The proposed SAPGAN-GPCA-WLP-CDC-SHTD method provides 76.92%, 42.86%, and 45.45% lower MSPE for prediction interval 20 min; 80.65%, 76.09%, and 78.95% lower MSPE for prediction interval 40 min; 66.67%, 69.23%, and 55.56% lower MSPE for prediction interval 60 min; 69.57%, 53.33%, and 41.67% lower MSPE for prediction interval 80 min; 65.52%, 47.06%, and 27.27% lower MSPE for prediction interval 100 min compared with existing methods, like AADEA-WLP-CDC-SHTD, BALNN-WLP-CDC-SHTD, and EMD-LSTM-

TABLE 3 Comparative table for Saskatchewan HTTP traces dataset.

Author name	Accuracy (%)	Precision (%)	Recall (%)	Energy consumption (J)	Correlation coefficient	Mean squared prediction error
Saxena and Singh[17]	72.55	69.35	74.58	0.96	0.48	0.38
Kumar et al.[18]	74.33	71.43	75.44	0.84	0.53	0.24
Yazdanian and Sharifian[19]	76.92	72.85	77.08	0.92	0.74	0.36
Al-Asaly et al.[20]	77.12	74.14	78.32	0.75	-	0.33
Khan et al.[21]	79.32	-	80.22	0.48	0.54	0.18
Singh et al.[22]	80.53	79.23	81.56	-	0.76	0.13
Saxena et al.[23]	81.33	-	83.34	0.43	0.67	0.26
Jeddi and Sharifian[24]	-	83.11	85.82	-	0.76	0.11
SAPGAN-GPCA-WLP-CDC-ND (proposed)	99.48	99.84	99.74	0.24	0.98	0.04

GAN-WLP-CDC-SHTD. The comparative table for NASA dataset and Saskatchewan HTTP traces Dataset is given in Tables 2 and 3.

5 | CONCLUSION

The SAPGAN using GPCA is successfully implemented in this manuscript for the prediction of workload as well as maximizing the efficiency of power in CDC. The data are gathered from the aforementioned datasets. The simulation of the proposed technique is activated in JAVA, and its efficacy is examined under the performance metrics. The proposed SAPGAN-GPCA-WLP-CDC method attains 27.5%, 10.32%, and 16.65% higher precision; 30.93%, 11.14%, and 15.3% higher recall for NASA dataset; and 36.31%, 15.78%, and 28.08% higher precision; 30.15%, 11.72%, and 18.34 higher recall for the dataset of Saskatchewan HTTP traces when analyzed to the existing AADEA-WLP-CDC, BALNN-WLP-CDC, and EMD-LSTM-GAN-WLP-CDC methods. In the future, we intend to create a resource allocation approach that is effective by utilizing the predicted load from SAPGAN to eliminate spare resources. It will reduce the cost of bandwidth, power consumption, and SLA violations. We also intend to evaluate the performance of SAPGAN in different time series prediction applications.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

G. Saravanan https://orcid.org/0000-0001-8403-6606

REFERENCES

1. Kumar J, Singh AK, Buyya R. Self-directed learning based workload forecasting model for cloud resource management. *Inform Sci*. 2021;543:345-366. doi:10.1016/j.ins.2020.07.012
2. Zakeri F, Mariethoz G. A review of geostatistical simulation models applied to satellite remote sensing: methods and applications. *Remote Sens Environ*. 2021;259:112381. doi:10.1016/j.rse.2021.112381
3. Bi J, Li S, Yuan H, Zhou M. Integrated deep learning method for workload and resource prediction in cloud systems. *Neurocomputing*. 2021;424:35-48. doi:10.1016/j.neucom.2020.11.011
4. Abdelhalim IS, Mohamed MF, Mahdy YB. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Syst Appl*. 2021;165:113922. doi:10.1016/j.eswa.2020.113922
5. Ouhame S, Hadi Y, Ullah A. An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model. *Neural Comput Applic*. 2021;33(16):10043-10055. doi:10.1007/s00521-021-05770-9
6. Shahidinejad A, Ghobaei-Arani M, Masdari M. Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. *Clust Comput*. 2021;24(1):319-342. doi:10.1007/s10586-020-03107-0

7. Praveenchandar J, Tamilarasi A. Dynamic resource allocation with optimized task scheduling and improved power management in cloud computing. *J Ambient Intell Humaniz Comput*. 2021;12(3):4147-4159. doi:10.1007/s12652-020-01794-6
8. Ruan L, Bai Y, Li S, He S, Xiao L. Workload time series prediction in storage systems: a deep learning based approach. *Clust Comput*. 2021;1-1(1):25-35. doi:10.1007/s10586-020-03214-y
9. Ghobaei-Arani M, Shahidinejad A. An efficient resource provisioning approach for analyzing cloud workloads: a metaheuristic-based clustering approach. *J Supercomput*. 2021;77(1):711-750. doi:10.1007/s11227-020-03296-w
10. Pushpalatha R, Ramesh B. Amalgamation of neural network and genetic algorithm for efficient workload prediction in data center. In: *Advances in VLSI, signal processing, power electronics, IoT, communication and embedded systems: select proceedings of VSPICE 2020*. Springer; 2021:69-84.
11. Masdari M, Khezri H. Efficient VM migrations using forecasting techniques in cloud computing: a comprehensive review. *Clust Comput*. 2020;23(4):2629-2658. doi:10.1007/s10586-019-03032-x
12. Masdari M, Zangakani M. Green cloud computing using proactive virtual machine placement: challenges and issues. *J Grid Comput*. 2020;18(4):727-759. doi:10.1007/s10723-019-09489-9
13. Harifi S, Mohammadzadeh J, Khalilian M, Ebrahimnejad S. Giza pyramids construction: an ancient-inspired metaheuristic algorithm for optimization. *Evol Intell*. 2021;14(4):1743-1761. doi:10.1007/s12065-020-00451-3
14. Ilager S, Ramamohanarao K, Buyya R. Thermal prediction for efficient energy management of clouds using machine learning. *IEEE Trans Parallel Distrib Syst*. 2020;32(5):1044-1056. doi:10.1109/TPDS.2020.3040800
15. Kalyampudi PL, Krishna PV, Kuppani S, Saritha V. A work load prediction strategy for power optimization on cloud based data centre using deep machine learning. *Evol Intell*. 2021;14(2):519-527. doi:10.1007/s12065-019-00289-4
16. Kumar J, Singh AK. Performance evaluation of metaheuristics algorithms for workload prediction in cloud environment. *Appl Soft Comput*. 2021;113:107895. doi:10.1016/j.asoc.2021.107895
17. Saxena D, Singh AK. Auto-adaptive learning-based workload forecasting in dynamic cloud environment. *Int J Comput Appl*. 2022;44(6):541-551. doi:10.1080/1206212X.2020.1830245
18. Kumar J, Saxena D, Singh AK, Mohan A. Biphase adaptive learning-based neural network model for cloud datacenter workload forecasting. *Soft Comput*. 2020;24(19):14593-14610. doi:10.1007/s00500-020-04808-9
19. Yazdanian P, Sharifian S. E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction. *J Supercomput*. 2021;77(10):11052-11082. doi:10.1007/s11227-021-03723-6
20. Al-Asaly MS, Bencherif MA, Alsanad A, Hassan MM. A deep learning-based resource usage prediction model for resource provisioning in an autonomic cloud computing environment. *Neural Comput Applic*. 2021;1-8(13):10211-10228. doi:10.1007/s00521-021-06665-5
21. Khan T, Tian W, Ilager S, Buyya R. Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. *Future Gener Comput Syst*. 2022;128:320-332. doi:10.1016/j.future.2021.10.019
22. Singh AK, Saxena D, Kumar J, Gupta V. A quantum approach towards the adaptive prediction of cloud workloads. *IEEE Trans Parallel Distrib Syst*. 2021;32(12):2893-2905. doi:10.1109/TPDS.2021.3079341
23. Saxena D, Singh AK, Buyya R. OP-MLB: an online VM prediction-based multi-objective load balancing framework for resource management at Cloud Data Center. *IEEE Trans Cloud Comput*. 2021;10(4):2804-2816. doi:10.1109/TCC.2021.3059096
24. Jeddi S, Sharifian S. A hybrid wavelet decomposer and GMDH-ELM ensemble model for network function virtualization workload forecasting in cloud computing. *Appl Soft Comput*. 2020;88:105940. doi:10.1016/j.asoc.2019.105940

How to cite this article: Saravanan G, Santhosh Babu AV. Workload prediction for enhancing power efficiency of cloud data centers using optimized self-attention-based progressive generative adversarial network. *Int J Commun Syst*. 2023;e5634. doi:10.1002/dac.5634