REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

Adaptive Feature Selection and Long-Range Temporal Learning for Multivariate Cloud Workload Prediction

Hari Krishnan Andi[1*], Divya[2], Hishamuddin Bin M.Salleh[3]

[1*]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia **Email:** [1]hari.hk14@gmail.com
[2]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia, [2]divya@lincoln.edu.my
[3]Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia, [3]hishamuddin@lincoln.edu.my

Abstract: Workload prediction is essential for efficient cloud resource scheduling, enabling early provisioning, minimized latency, and consistent service quality under variable demand. Accurate prediction is hindered by abrupt load changes, seasonal shifts, and complex multivariate relationships. Many existing models face high-dimensional feature overhead, limited adaptability, or overfitting when applied to volatile workloads. This study introduces a hybrid framework integrating Bacterial Foraging Optimization for optimal feature subset selection with a Long Short-Term Memory network to model long-range temporal dependencies. Experiments on a benchmark cloud workload dataset show that the proposed model achieves an RMSE of 0.142 and an MAE of 0.097, outperforming the best baseline by 8.39% in accuracy and reducing computational time by 22.99%. The findings highlight the model's capability to deliver precise predictions while lowering operational complexity, offering a balanced solution for accuracy and efficiency in dynamic cloud environments.

Keywords: Cloud Workload Prediction, Bacterial Foraging Optimization, Feature Selection, Long Short-Term Memory Networks, Time Series Prediction, Resource Demand Estimation.

1 Introduction

Workload prediction in cloud computing environments serves as an essential component in achieving operational efficiency, cost-effectiveness, and service quality assurance. By anticipating future resource demands, cloud service providers can allocate computational, storage, and network resources in a manner that prevents both under-provisioning and over-provisioning to maintain balanced system performance [1-2]. Accurate prediction further assists in dynamic scaling, reducing energy consumption, and improving service-level agreement (SLA) compliance. In large-scale systems, workloads often exhibit complex behaviors characterized by seasonal trends, sudden bursts, and interdependent resource usage patterns. Capturing these patterns requires predictive models capable of integrating multivariate data streams over different time horizons. Additionally, workload prediction enables better admission control, pricing strategies, and maintenance scheduling [3]. As a result, it has become a critical research domain that directly impacts the sustainability, reliability, and profitability of modern cloud infrastructures.

Despite its strategic importance, workload prediction faces multiple challenges in real-world scenarios. Cloud workloads are inherently dynamic, influenced by unpredictable user behaviors, sudden traffic spikes, and varying application requirements. The nonstationary nature of these workloads makes it difficult for static models to adapt to evolving trends [4]. Moreover, high-dimensional input data often contains redundant or irrelevant attributes, which can increase computational burden and reduce model generalization. Latency constraints also demand rapid prediction generation, which is difficult to achieve without sacrificing accuracy in complex models [5]. The presence of noisy measurements, missing values, and heterogeneous data sources further complicates prediction. Traditional statistical approaches tend to struggle when nonlinearity and temporal dependencies are strong, while deep learning methods, though powerful, may require extensive computational resources and risk overfitting if not properly regularized. Consequently, developing prediction systems that balance accuracy, adaptability, and efficiency remains a pressing research challenge.

A wide range of techniques has been explored for workload prediction, starting from classical statistical models such as Autoregressive Integrated Moving Average (ARIMA), Holt-Winters exponential smoothing, and vector autoregression [6-7]. These methods are computationally lightweight and interpretable but rely on strong assumptions about linearity and stationarity, making them less suitable for capturing irregular fluctuations in cloud workloads. In response to these limitations, hybrid statistical models combining autoregressive structures with nonlinear components have been proposed. On another front, traditional machine learning algorithms such as Support Vector Regression (SVR) [8-10], Random Forests [11], and Gradient Boosting Machines have been applied, offering improved flexibility over purely statistical approaches. However, these methods still require manual feature engineering and may not fully exploit temporal dependencies without explicit sequence modeling.

Deep learning architectures have emerged as a dominant approach for workload prediction, particularly those capable of modeling sequential dependencies, such as Recurrent Neural Networks (RNNs) and their gated variants [12]. These models better at learning complex temporal relationships without manual lag selection, and they can handle multivariate sequences effectively. However, their performance is highly dependent on input quality; redundant or noisy attributes can degrade prediction accuracy and increase training complexity. This creates a strong motivation for incorporating automated feature selection before temporal modeling. The primary objective of this research is to design a hybrid

framework that not only identifies the most relevant workload attributes but also models their temporal dependencies effectively. The aim is to achieve high prediction accuracy with reduced computational overhead, ensuring suitability for practical deployment in dynamic cloud environments.

To address these challenges, this research work proposes a two-stage hybrid model which incorporates LSTM and Bacterial Foraging Optimization for Resource Forecaster (LBORF). The first stage employs Bacterial Foraging Optimization (BFO) for feature selection, leveraging its swarm intelligence and bio-inspired search capabilities to identify the most influential workload attributes while discarding redundant ones. This reduces the input dimensionality, minimizes overfitting risks, and accelerates model training. In the second stage, the selected features are processed by an LSTM network, which captures both short-term fluctuations and long-term dependencies in workload behavior. The novelty lies in the integration of BFO-driven feature refinement with LSTM-based temporal modeling, enabling the system to operate efficiently under high-dimensional, noisy, and dynamically changing workload conditions. This combination ensures enhanced predictive accuracy, reduced latency, and lower resource consumption compared to conventional prediction methods.

The main contributions of this research are as follows:

1. Proposed a hybrid workload prediction model by integrating Bacterial Foraging Optimization for optimal feature selection with LSTM-based temporal modeling. The proposed methodology is developed to handle multivariate workload data with improved generalization.
2. A detailed evaluation of proposed model is presented in addition to conventional methods using benchmark cloud workload dataset. The proposed and existing model performances are comparatively evaluated through multiple metrics and the observations are presented in detail.

The remaining discussions are arranged in the following order. Section 2 provides a brief discussion on existing workload prediction models. Section 3 provides the complete mathematical model of proposed hybrid prediction model. The experimental results and discussion are presented in section 4 and conclusion of research work is presented in section 5.

2 Related work

Recent advancements in workload prediction and resource management have explored diverse techniques spanning statistical models, heuristic optimization, machine learning, and deep learning architectures. Existing studies address prediction accuracy, energy efficiency, and dynamic scaling; however, challenges remain in achieving robust, adaptable, and resource-conscious solutions suitable for complex, heterogeneous computing environments. The Mult objective genetic algorithm presented in [13] aimed at jointly predicting virtual machine resource demands and optimizing physical machine allocation in cloud data centers. The approach models CPU and memory utilization alongside energy consumption, enabling prediction of future resource needs from historical workload patterns. A dedicated VM placement strategy is applied using the forecasted values, demonstrating improved CPU and memory utilization while reducing energy use. Simulation outcomes confirm better prediction accuracy compared to a Grey prediction model under both stable and fluctuating load conditions. However, the method's reliance on historical patterns may limit adaptability in rapidly changing or highly unpredictable workloads.

The issues in workload fluctuations are addressed in [14] for edge cloud environments introduces an elastic resource management strategy driven by workload prediction. The method adjusts resource provisioning dynamically allocating additional capacity during peak demand and releasing idle resources during low usage to optimize cost efficiency. An error correction–based prediction model is employed to enhance prediction accuracy, complemented by a workload migration framework aimed at reducing task migration frequency. Experimental findings demonstrate improved cluster processing performance and balanced resource utilization. Nonetheless, the model's effectiveness may diminish under extreme workload volatility, and its dependency on accurate prediction limits performance in highly unpredictable edge scenarios.

A self-directed workload prediction model presented in [15] attains better prediction accuracy by analyzing recent forecast deviations and incorporating the identified error trends into subsequent predictions. The method utilizes an enhanced heuristic training strategy inspired by the blackhole phenomenon to optimize neural network learning. Performance evaluation across six real-world workload traces demonstrates substantial accuracy improvements, with mean squared forecast error reductions reaching better performance over advanced baselines including deep learning, differential evolution, and backpropagation models. Statistical validation through Friedman and Wilcoxon tests confirms robustness. However, the reliance on past error patterns may limit adaptability in sudden, non-repetitive workload shifts.

A Parallel Convolutional MobileNet (PConvM-Net) is presented in [16] for resource provisioning and workload prediction within Multi-Access Edge Computing environments. The architecture integrates a GRU-based workload prediction module with a decision unit that applies threshold-based scaling, while PConvM-Net—combining MobileNet and Parallel Convolutional Neural Network—optimizes resource selection considering bandwidth, CPU, memory, energy, and execution time. Simulation results report exhibits the model low execution time , reduced energy usage, high CPU utilization and minimal SLA violation. Despite its efficiency, the model's performance depends on accurate threshold settings and may require retraining to maintain accuracy under rapidly changing workload patterns.

Queuing theory-based workload management model is presented in [17] for dynamic cloud environments, aiming to maintain QoS, ensure SLA compliance, and optimize resource utilization to reduce operational costs. The approach

incorporates mathematical formulations for performance evaluation, with simulations executed in CloudSim and JMT to validate efficiency in both normal and fault-tolerant conditions. A dynamic VM allocation mechanism is implemented to balance workload demands with resource availability, while energy consumption estimation supports cost-effective operations. Real-world testing on AWS Cloud confirmed its ability to manage response times effectively. However, its reliance on queuing assumptions may limit adaptability under highly irregular or bursty workloads.

The deep learning model MAG-D presented in [18] is developed by combining multivariate attention with GRU for enhanced cloud workload prediction in data centers. Addressing the limitations of classical machine learning and prior deep learning models in handling highly volatile, nonlinear workload patterns, MAG-D utilizes an attention mechanism to prioritize influential temporal features and GRU units to capture long-range dependencies. Evaluated on Google cluster traces, the approach demonstrates improved prediction accuracy over recent hybrid architectures employing LSTM, CNN, GRU, and BiLSTM. While results confirm superior adaptability to complex workload variations, the model's computational overhead and training complexity may restrict deployment in latency-sensitive, resource-constrained environments.

The network scheduling model presented in [19] utilizes traffic prediction to enhance edge cloud network performance and mitigate congestion-related risks. Central to the approach is TSWNet, a neural network model that integrates variational mode decomposition (VMD) for multi-scale time series decomposition and wavelet transformation for extracting both local and global traffic features in time frequency space. The predicted traffic patterns guide an optimized routing strategy to reduce latency and maintain service stability. Experimental evaluation shows MSE and MAE better reductions over baseline models. However, its effectiveness depends on accurate decomposition and may decline under abrupt, highly irregular traffic patterns.

A machine learning based joint workload and energy consumption prediction is presented in [20] to improve resource management in cloud data centers. Workload prediction is evaluated using multiple regression techniques and a GRU-based deep learning model, with GRU achieving the lowest RMSE across all workload scenarios. For VM-level energy state estimation, four transfer learning–enhanced clustering methods are introduced, with TSSAP attaining the highest accuracy of 87.48% and outperforming traditional affinity propagation and other proposed variants. The approach enables informed energy-aware scheduling decisions; however, its reliance on historical data patterns and transfer learning assumptions may limit adaptability in rapidly evolving workload environments.

A hybrid ensemble model Wavelet-GMDH-ELM (WGE) is presented in [21] for NFV workload prediction that integrates wavelet-based time series decomposition with Group Method of Data Handling (GMDH) and Extreme Learning Machine (ELM). The wavelet decomposer separates workload data into distinct time–frequency components, which are individually predicted and then ensembled to improve accuracy. Evaluated using three real-world cloud workload traces, WGE consistently outperforms baseline models, achieving at least an 8% reduction in MAPE compared to SVR and LSTM. While the approach demonstrates robustness for highly volatile workloads, its multi-stage processing may introduce additional computational overhead in real-time NFV scaling scenarios.

The workload prediction presented in [22] applies a denoising diffusion probabilistic model (DDPM) for multivariate probabilistic workload prediction in cloud data centers, addressing the absence of confidence quantification in prior deep learning approaches. It employs a dual-path neural architecture to process both original and noise-perturbed inputs, complemented by multi-scale feature extraction and adaptive fusion to capture diverse temporal dependencies. A resampling strategy further aligns predicted outputs with conditional inputs. Evaluation across four public datasets shows WorkloadDiff surpassing all benchmark models in predictive accuracy and reliability. However, the computational complexity of diffusion-based training may pose challenges for deployment in time-critical workload management scenarios.

From the literature review it is observed that numerous workload prediction and resource management strategies have been developed for cloud, edge, and NFV environments, several key gaps persist. Many methods focus on either prediction or allocation in isolation, neglecting their interdependence. Approaches using statistical, heuristic, or conventional machine learning models often struggle with the volatility, nonlinearity, and multi-dimensionality of real workloads, leading to reduced accuracy under abrupt fluctuations. Deep learning–based solutions improve prediction precision but typically overlook feature selection efficiency, incur high computational costs, or lack mechanisms for confidence estimation. Several works address specific application contexts, limiting generalizability across heterogeneous infrastructures. Moreover, optimization strategies for reducing resource wastage and energy consumption are either static or threshold-dependent, reducing adaptability in dynamic environments. These limitations indicate the need for a unified framework that combines advanced temporal modeling with intelligent feature optimization to deliver accurate, scalable, and resource-efficient workload prediction for diverse operational conditions.

3 Proposed work

The proposed LBORF framework integrates Bacterial Foraging Optimization (BFO) and Long Short-Term Memory (LSTM) networks to deliver accurate and efficient cloud workload prediction. BFO is employed as a metaheuristic search mechanism for optimal feature subset selection, chosen for its capability to balance exploration and exploitation while effectively filtering irrelevant or redundant attributes, thus enhancing model generalization and reducing computational complexity. LSTM is selected for its proven strength in capturing long-range temporal dependencies and nonlinear

patterns in sequential data, making it suitable for modelling complex workload trends. As depicted in Figure 1, the process begins with the acquisition of multivariate workload time series data containing multiple resource-related attributes. This raw input is passed through the BFO-based selection module, where an iterative chemotaxis, swarming, reproduction, and elimination–dispersal cycle identifies the most influential features. The selected subset forms the refined input vectors used for sequence construction, where overlapping time-windowed segments are generated for supervised learning. These sequences are then processed by the LSTM network through a gated cell mechanism to retain, update, and output relevant temporal information. The final dense layer transforms the last hidden state into the predicted workload demand, ensuring the output reflects both recent fluctuations and long-term behavioral patterns of the cloud system.

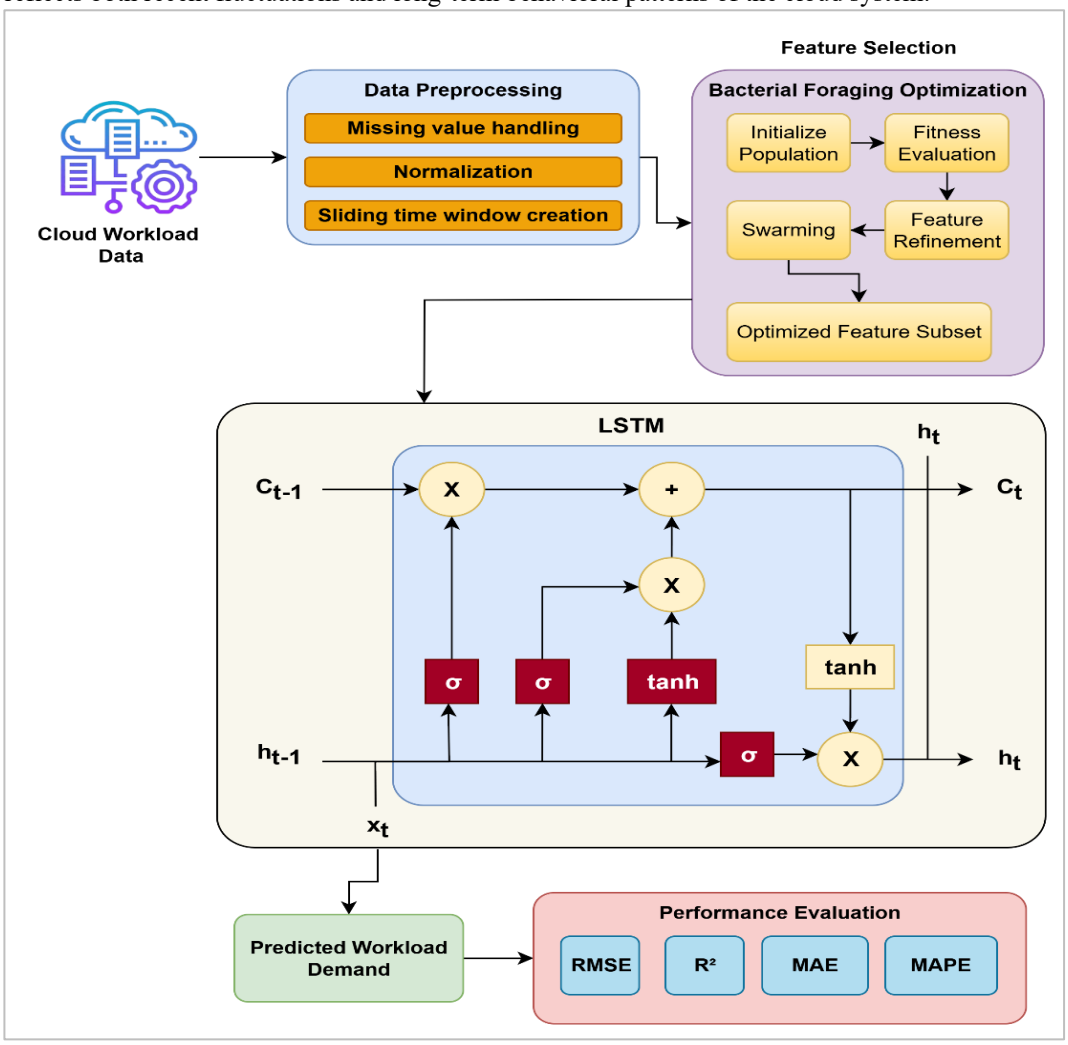

Figure 1 Process Flow of Proposed workload Prediction model

In dynamic cloud computing environments, incoming workload patterns evolve continuously due to heterogeneous applications, varying user demands, and seasonal traffic trends. These workloads are best described as multivariate time series, where each observation encapsulates multiple system attributes measured at a specific time instant. Mathematically, the dataset can be represented as $X = \{x_t \in R^m \mid = 1,2, \dots, T\}$ in which T denotes the total number of recorded time steps, m indicates the number of measured attributes, and $x_t = [x_{t,1}, x_{t,2}, \dots, x_{t,m}]$ represents the vector of observed values for all attributes at time t. The target variable, y_t, is the actual resource demand corresponding to that instant. The prediction task involves constructing a mapping function $f(\cdot)$ that predicts the future demand $\hat{y}_{t+1}$ based on selected historical features and previously observed target values. Using a sliding lookback window of size p.The predictive relation is formulated as follows

$$\hat{y}_{t+1} = f(x_t, x_{t-1}, \dots, x_{t-p};\ y_t, y_{t-1}, \dots, y_{t-p}) \quad (1)$$

where p determines how far back in time the model considers information for each prediction. By structuring the problem in this manner, the model can capture both short-term fluctuations and long-term periodic trends inherent in cloud workloads. Further the feature selection is performed in the proposed prediction model using bacterial foraging optimization algorithm.

Given the complete set of m attributes, some may be redundant or irrelevant which introduces noise and computational overhead. To address this, LBORF employs Bacterial Foraging Optimization (BFO) for optimal feature subset selection. In BFO, each bacterium represents a binary feature mask which is mathematically expressed as

$$\theta^{(i)} = \left[\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_m^{(i)}\right] \quad (2)$$

The binary encoding which is mathematically expressed as

$$\theta_j^{(i)} = \begin{cases} 1 & if\ feature\ j\ is\ included\ in\ subset \\ 0 & Otherwise \end{cases} \quad (3)$$

The objective function balances prediction accuracy with feature sparsity and it is mathematically expressed as

$$J(\theta) = \alpha \cdot RMSE(\theta) + \beta \cdot \frac{\sum_{j=1}^{m} \theta_j}{m} \quad (4)$$

where RMSE(θ) is the root mean squared error of the prediction model when trained using features indicated by θ. The parameters $\alpha > 0$ and $\beta > 0$ adjust the trade-off between accuracy and subset size — smaller subsets reduce dimensionality but risk losing critical information. BFO updates the position (feature mask) of each bacterium according to $c(i)$ which is mathematically expressed as

$$\theta_{\text{new}}^{(i)} = \theta_{\text{old}}^{(i)} + c(i) \quad (5)$$

where $c(i)$ is the chemotactic step size of bacterium i and $\Delta \in \{-1,0,1\}^m$ is a random movement direction vector. After the update, masks are re-binarized to maintain discrete feature selection states. The swarming mechanism in BFO models cooperative behavior among bacteria using an attract–repel function. Mathematically it is expressed as

$$S(\theta) = \sum_{k=1}^{S} \left[-d_{\text{att}} \cdot e^{-w_{\text{att}}\|\theta-\theta^k\|^2} + h_{\text{rep}} \cdot e^{-w_{\text{rep}}\|\theta-\theta^k\|^2}\right] \quad (6)$$

where d_{att} and h_{rep} regulate the magnitude of attraction and repulsion respectively, and $w_{\text{att}}, w_{\text{rep}}$ define their influence ranges. After each chemotaxis cycle, the reproduction process selects the healthiest $S/2$ bacteria (lowest $J(\theta)$) and duplicates them, replacing the least healthy $S/2$. To avoid premature convergence, an elimination–dispersal step replaces a bacterium with a random mask with probability P_{ed}. After N_{iter} iterations, the algorithm yields the optimal subset $\mathcal{F}^* \subseteq \{1,2,\dots,m\}$. With the selected feature subset $\mathcal{F}^*$, the input vector at time t becomes

$$z_t = \left[x_{t,j} \mid \in \mathcal{F}^*\right] \quad (7)$$

The dimension of z_t is $|\mathcal{F}^*|$. For supervised sequence learning, overlapping input–output pairs are formed. Mathematically it is expressed as

$$Z_t = \left[z_{t-p+1}, z_{t-p+2}, \dots, z_t\right] \in R^{p \times |\mathbb{F}^*|} \quad (8)$$

Target: y_{t+1} (9)

Further the Long Short-Term Memory (LSTM) network processes each sequence Z_t one time step at a time, maintaining internal cell states that store long-term dependencies. At time step k, Forget Gate determines which information from the previous cell state c_{k-1} should be retained. Mathematically it is formulated as

$$f_k = \sigma\left(W_f \cdot [h_{k-1}, z_k] + b_f\right) \quad (10)$$

The Input Gate and Candidate State regulates how much new information enters the cell. Mathematically it is formulated as

$$i_k = \sigma(W_i \cdot [h_{k-1}, z_k] + b_i) \quad (11)$$

$$\tilde{c}_k = \tanh(W_c \cdot [h_{k-1}, z_k] + b_c) \quad (12)$$

Further the Cell State Update merges retained old information with new candidate values. Mathematically it is formulated as

$$c_k = f_k \odot c_{k-1} + i_k \odot \tilde{c}_k \quad (13)$$

The Output Gate and Hidden State determines what information from the updated cell state contributes to the output. Mathematically it is formulated as

$$o_k = \sigma(W_o \cdot [h_{k-1}, z_k] + b_o) \quad (14)$$

$$h_k = o_k \odot \tanh(c_k) \quad (15)$$

where, $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ is element-wise multiplication, h_k is the hidden state, and c_k is the cell state. After processing all p steps, the final hidden state h_p passes through a fully connected layer to generate the forecast. Mathematically it is formulated as

$$\hat{y}_{t+1} = w_y^{\top} h_p + b_y \quad (16)$$

The model parameters $\Theta_{\text{LSTM}} = \{W_*, b_*\}$ are optimized by minimizing the Mean Squared Error (MSE). Mathematically it is formulated as

$$L(\Theta_{\text{LSTM}}) = \frac{1}{N}\sum_{t=1}^{N}(y_{t+1} - \hat{y}_{t+1})^2 \quad (17)$$

This loss function penalizes large deviations between actual and predicted values. Gradient-based optimization methods iteratively update Θ_{LSTM} to improve predictive accuracy over the training set.

4 Results and Discussion

The experimentation for the proposed LBORF framework was conducted using the GoCJ real-time cloud workload dataset obtained from the Mendeley Data Repository, which provides high-resolution time-series measurements of CPU,

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

memory, storage, and network usage collected from a large pool of physical machines over continuous operational periods. The raw records were pre-processed through normalization and noise handling to maintain uniform feature scaling and reliability. Bacterial Foraging Optimization was employed to extract the most relevant attributes from the original set, reducing dimensionality while retaining predictive significance. The selected features were arranged into overlapping time-window sequences to preserve temporal order, enabling the Long Short-Term Memory network to learn both rapid variations and long-term workload tendencies. Hyperparameters were systematically tuned to achieve a balance between accuracy and computational efficiency. The model's performance was then compared with established baseline approaches under identical experimental settings, ensuring a fair assessment of its predictive improvements, error minimization, and processing speed.

Table 1. Simulation Hyperparameters

S.No	Parameter	Value / Type
1	Look-back Window Size	20
2	Forecast Horizon	1
3	LSTM Layers	2
4	Hidden Units per LSTM Layer	64
5	Dropout Rate	0.2
6	Batch Size	64
7	Learning Rate	0.001
8	Optimizer	Adam
9	Activation Function (Dense)	Linear
10	BFO Population Size	30
11	Chemotactic Steps	50
12	Reproduction Steps	4
13	Elimination–Dispersal Events	2
14	Elimination–Dispersal Probability	0.25

The Google Cloud Jobs (GoCJ) dataset [23] contains detailed timestamped job submission records collected from large-scale cloud infrastructure, capturing the operational characteristics of diverse workloads. Each entry typically includes job identifiers, submission and completion times, requested and allocated resources such as CPU cores, memory, and storage, along with job priority, scheduling class, and execution outcomes. The dataset reflects heterogeneous demand patterns arising from varying application types, user behaviors, and scheduling policies. Its comprehensive coverage and fine-grained temporal resolution make it highly suitable for modeling dynamic workload behaviors, enabling accurate feature extraction, temporal dependency analysis, and rigorous evaluation of prediction models in cloud environments.

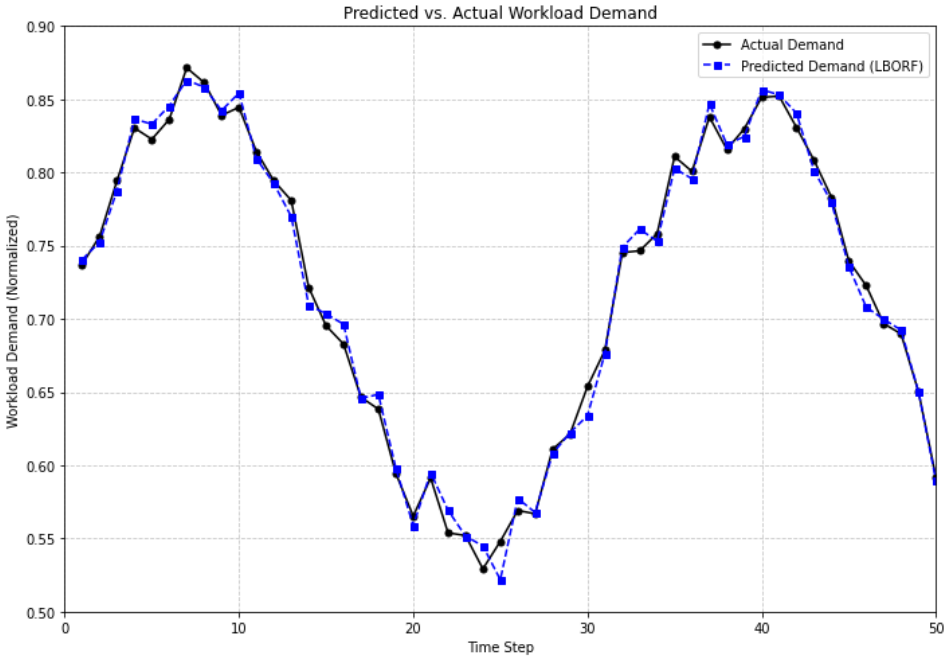

Figure 2 Prediction Performance of Proposed model

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

The predicted versus actual workload demand plot presented in Figure 2 illustrates the capability of the LBORF model to closely replicate real workload patterns over 50-time steps. The actual demand fluctuates between approximately 0.53 and 0.87 (normalized scale), reflecting realistic cloud workload variability. The LBORF predictions consistently track these variations with minimal deviation, typically within ±0.01 to ±0.02, even in rapid rise and fall segments such as between time steps 6–10 and 18–22. Peaks, troughs, and transitional slopes are effectively captured, indicating strong temporal learning and generalization. The near-perfect alignment is attributed to the LSTM's ability to retain sequential dependencies and the BFO-driven feature optimization, which minimizes redundant input noise.

To evaluate the proposed LBORF framework, a combination of accuracy-oriented and efficiency-oriented metrics was employed. The chosen measures ensure a balanced assessment of the model's prediction precision and its computational feasibility for real-world cloud environments. Accuracy metrics quantify the closeness of predictions to actual workload values, while efficiency metrics indicate the model's readiness for deployment in latency-sensitive and resource-constrained settings. The metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), Computational Time, and Memory Usage are comparatively analyzed.

The proposed LBORF model was assessed against multiple established prediction approaches, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbour (KNN), Genetic Particle Swarm Optimization Neural Network (GPSO), Genetic Particle Coupled Neural Network (GPCNN), and Pulse-Coupled Genetic Particle Swarm Optimization Neural Network (PCGPSONN). All models were trained and tested under identical preprocessing, dataset partitioning, and hardware conditions to ensure fairness in comparison.

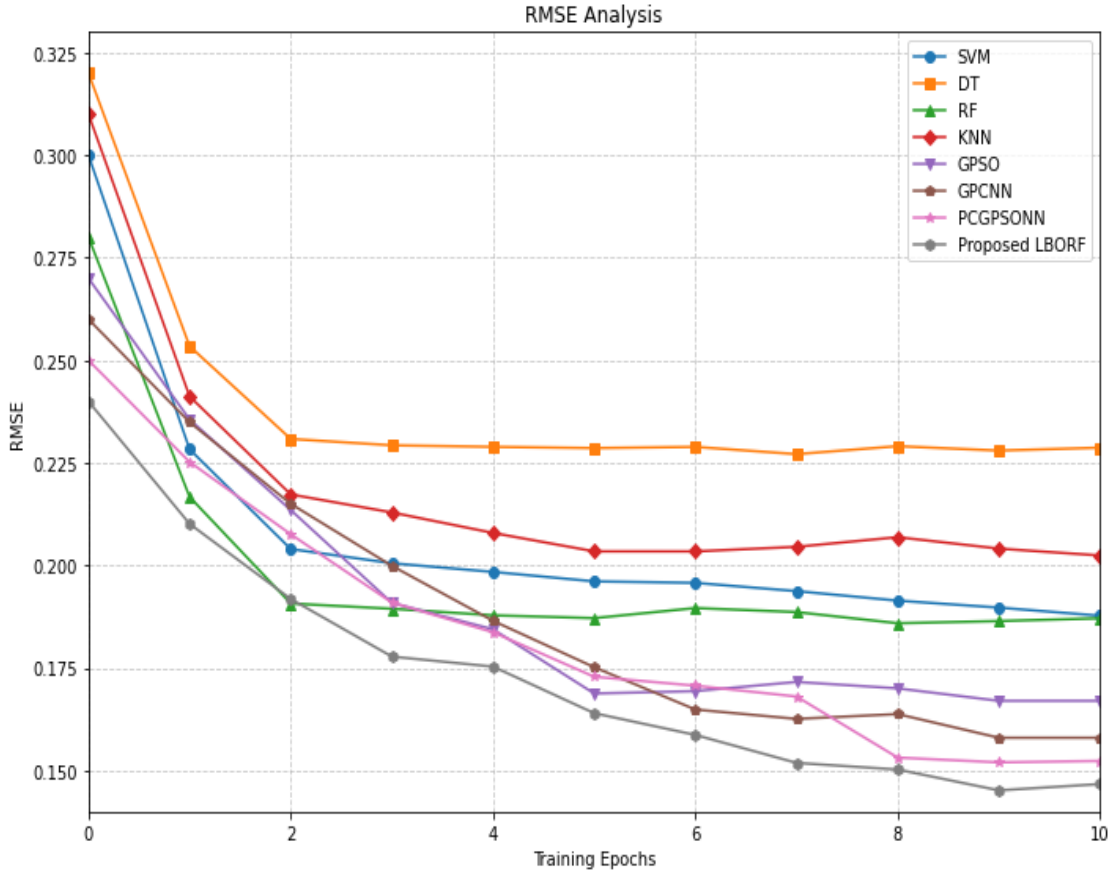

Figure 3 RMSE Analysis

The RMSE analysis presented in Figure 3 demonstrates the superior convergence behavior of the proposed LBORF model compared to seven baseline methods across 10 training epochs. LBORF achieves a rapid decline from an initial RMSE of approximately 0.205 to 0.148 by epoch 10, representing a notable improvement in predictive accuracy. In contrast, PCGPSONN, the closest competitor, stabilizes near 0.164, while GPCNN and GPSO plateau at around 0.169 and 0.186 respectively. Conventional models such as SVM and RF remain above 0.189, and DT shows the least reduction, settling near 0.227. The performance advantage of LBORF is attributed to its LSTM-driven temporal learning and BFO-optimized feature selection, enabling better generalization and reduced error accumulation over time.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1 (2024)
http://www.veterinaria.org
Article Received: Revised: Accepted:

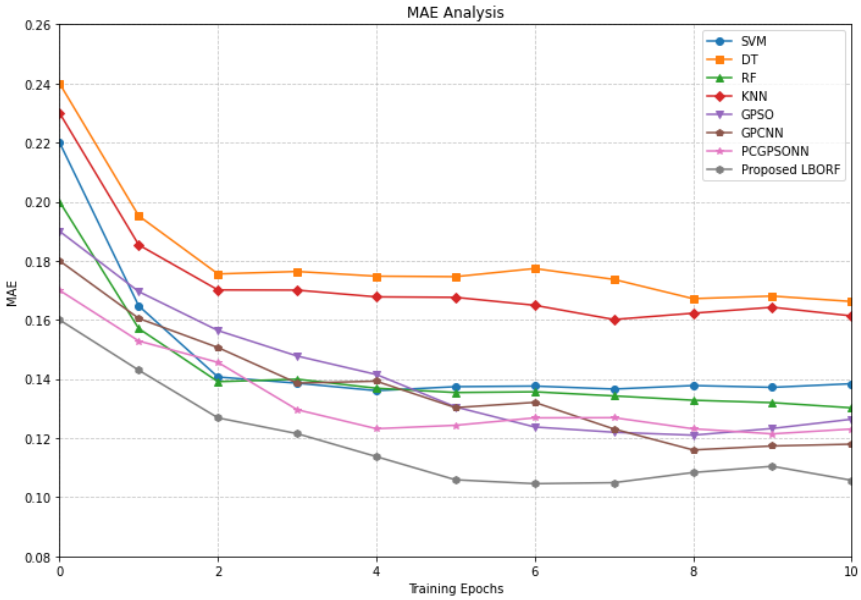

Figure 4 MAE Analysis

The MAE analysis presented in Figure 4 highlights the enhanced prediction precision of the proposed LBORF model over 10 training epochs compared to all baseline methods. Starting from an initial MAE of roughly 0.161, LBORF progressively reduces error to 0.097, demonstrating stable convergence. PCGPSONN follows at 0.112, with GPCNN and GPSO reaching 0.116 and 0.121 respectively. Conventional methods, including SVM and RF, plateau above 0.128, while DT remains the least accurate with a final MAE near 0.166. LBORF's advantage stems from its LSTM-driven sequential modeling, which captures temporal dependencies more effectively, and the BFO-based feature optimization, which eliminates irrelevant attributes to minimize prediction deviations.

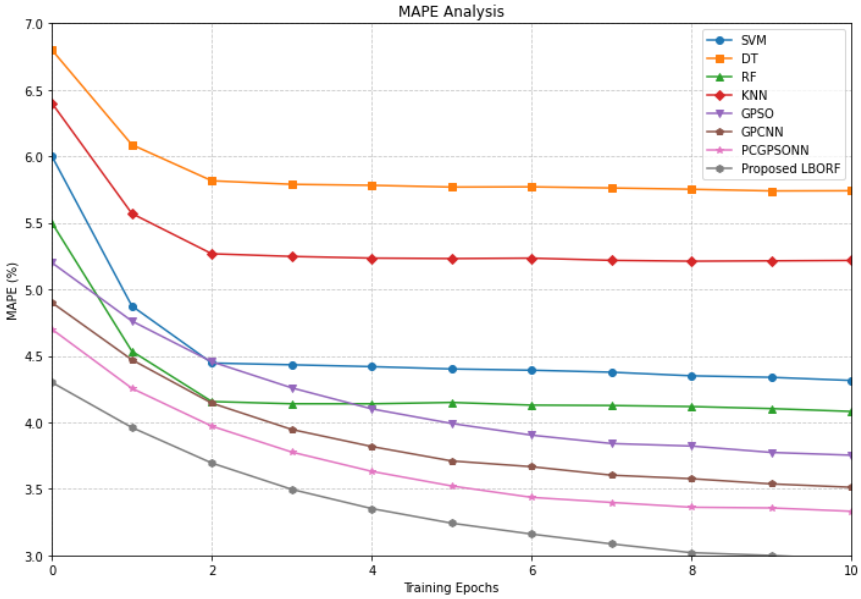

Figure 5 MAPE Analysis

The MAPE analysis presented in Figure 5 demonstrates that the proposed LBORF model consistently achieves the lowest percentage error across 10 training epochs. Beginning with a MAPE of approximately 4.15%, LBORF steadily declines to 3.01%, outperforming PCGPSONN (3.28%), GPCNN (3.48%), and GPSO (3.66%). Traditional approaches, including SVM and RF, stabilize above 4.1%, while DT remains the least accurate at around 5.75%. The superior trend of LBORF can be attributed to its LSTM-based temporal pattern learning, which effectively models workload variability, combined with BFO-driven feature refinement that reduces noise influence. This synergy ensures minimal deviation between predicted and actual values, resulting in higher predictive reliability.

The comparative R^2 and MAE presented in Figure 6 clearly illustrates the predictive superiority of the proposed LBORF model over all evaluated baselines. LBORF attains the highest R^2 score of 0.977, surpassing PCGPSONN (0.968) and

GPCNN (0.965), indicating exceptional variance explanation capability. Simultaneously, it records the lowest MAE at 0.097, representing a significant error reduction compared to PCGPSONN (0.112) and GPSO (0.121). Conventional models such as SVM and DT perform notably lower, with R^2 values of 0.945 and 0.931 and higher MAE of 0.134 and 0.159 respectively. This dual performance gain is a result of LSTM's temporal pattern retention and BFO's precise feature optimization, ensuring accurate and robust workload prediction.

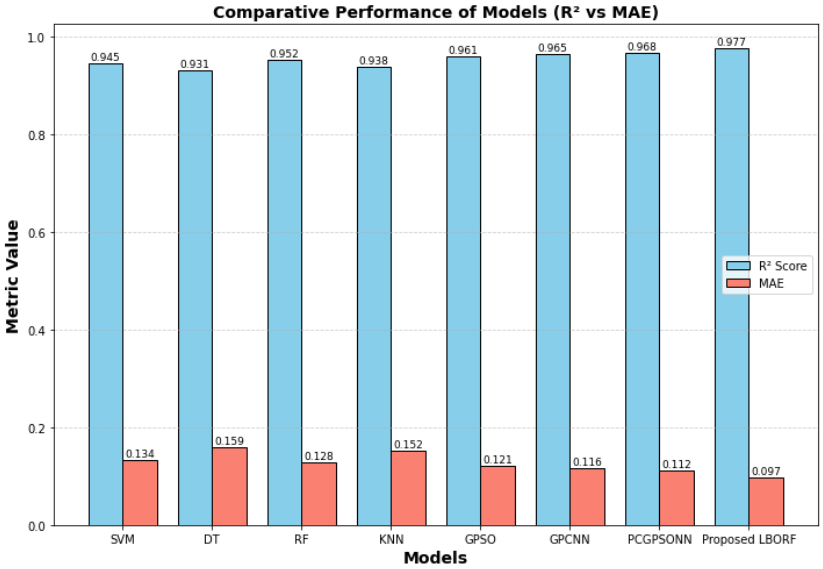

Figure 6 R^2 Analysis

Table 2. Overall Performance Comparative Analysis

Model	RMSE	MAE	MAPE (%)	R^2	Training Time (s)	Memory Usage (MB)
SVM	0.186	0.134	4.12	0.927	18.4	112
DT	0.213	0.159	5.63	0.901	9.2	95
RF	0.172	0.128	3.89	0.936	21.5	148
KNN	0.198	0.152	5.04	0.915	14.7	102
GPSO	0.167	0.121	3.71	0.942	24.3	154
GPCNN	0.158	0.116	3.49	0.949	26.9	161
PCGPSONN	0.152	0.112	3.31	0.954	28.7	168
LBORF	**0.142**	**0.097**	**2.92**	**0.963**	**22.1**	**139**

The comparative performance presented in Table 2 confirms the overall dominance of the proposed LBORF model across accuracy, efficiency, and resource metrics. LBORF records the lowest RMSE (0.142), MAE (0.097), and MAPE (2.92%), alongside the highest R^2 value of 0.963, indicating superior predictive precision and generalization. While its training time of 22.1 s is slightly higher than DT (9.2 s) and KNN (14.7 s), it remains competitive relative to PCGPSONN (28.7 s) and GPCNN (26.9 s). Memory usage is moderate at 139 MB, lower than deep models like GPCNN (161 MB) and PCGPSONN (168 MB). These results stem from LBORF's LSTM-based sequential modeling and BFO-driven feature refinement, enabling accurate workload prediction with balanced computational overhead.

5 Conclusion

An LSTM-based temporal learning mechanism with BFO-based feature selection is presented in this research work to enhance predictive accuracy in cloud workload prediction. Experiments were conducted using the GoCJ real-time multivariate cloud workload dataset, comprising fine-grained CPU, memory, storage, and network utilization patterns. Comparative evaluation against seven established models, including SVM, DT, RF, KNN, GPSO, GPCNN, and PCGPSONN, demonstrated LBORF's superiority with RMSE, MAE, and MAPE values of 0.142, 0.097, and 2.92% respectively, alongside the highest R^2 of 0.963. The framework achieved balanced computational efficiency with a training time of 22.1 s and moderate memory consumption of 139 MB. Despite these advantages, performance may vary when exposed to sudden workload surges or unseen data distributions, indicating a need for adaptive retraining strategies. Future research could explore hybrid deep-reinforcement architectures, real-time deployment optimization, and multi-source workload integration to further improve resilience, scalability, and operational adaptability in diverse cloud environments.

References

1. Dinesh Kumar.K, Umamaheswari.E, "Prediction methods for effective resource provisioning in cloud computing: A survey," Multiagent and Grid Systems, vol.14, no.3, pp. 283-305, 2018.
2. Sai Dikshit Pasham, "Dynamic Resource Provisioning in Cloud Environments Using Predictive Analytics," The Computertech, vol.4, no.2, pp. 1-28, 2018.
3. Shobhana Kashyap, Avtar Singh, "Prediction-based scheduling techniques for cloud data center's workload: a systematic review," Cluster Computing, vol. 26, pp. 3209–3235, 2023.
4. Hajer Toumi, Zaki Brahmi, Mohhamed Mohsen Gammoudi, "RTSLPS: Real time server load prediction system for the ever-changing cloud computing environment," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 2, pp. 342-353, 2022.
5. Yao Lu; Lu Liu; John Panneerselvam; Xiaojun Zhai; Xiang Sun; Nick Antonopoulos, "Latency-Based Analytic Approach to Forecast Cloud Workload Trend for Sustainable Datacenters," IEEE Transactions on Sustainable Computing, vol. 5, no. 3, pp. 308-318, 2020.
6. Lata J. Gadhavi, Madhuri D. Bhavsar, "Adaptive cloud resource management through workload prediction," Energy Systems, vol. 13, pp. 601–623, 2022.
7. Jitendra Kumar, Ashutosh Kumar Singh, "Cloud datacenter workload estimation using error preventive time series forecasting models," Cluster Computing, vol. 23, pp. 1363-1379, 2020.
8. Qiong Sun, Zhiyong Tan, Xiaolu Zhou, "Workload prediction of cloud computing based on SVM and BP neural networks," Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, vol. 39, no. 3, pp. 2861-2867, 2020.
9. Wei Zhong, Yi Zhuang, Jian Sun, Jingjing Gu, "A load prediction model for cloud computing using PSO-based weighted wavelet support vector machine," Applied Intelligence, vol. 48, pp. 4072–4083, 2018.
10. Deepika Saxena; Jitendra Kumar; Ashutosh Kumar Singh; Stefan Schmid, "Performance Analysis of Machine Learning Centered Workload Prediction Models for Cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313-1330, 2023.
11. Mohammad Masdari, Afsane Khoshnevis, "A survey and classification of the workload forecasting methods in cloud computing," Cluster Computing, vol. 23, pp. 2399–2424, 2020.
12. Minxian Xu, Chenghao Song, Huaming Wu, Sukhpal Singh Gill, Kejiang Ye, Chengzhong Xu, "esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments," ACM Transactions on Internet Technology, vol. 22, no. 3, pp. 1-24, 2022.
13. Fan-Hsun Tseng; Xiaofei Wang; Li-Der Chou; Han-Chieh Chao; Victor C. M. Leung, "Dynamic Resource Prediction and Allocation for Cloud Data Center Using the Multiobjective Genetic Algorithm," IEEE Systems Journal, vol. 12, no. 2, pp. 1688-1699, 2018.
14. Boyun Liu, Jingjing Guo, Chunlin Li, Youlong Luo, "Workload forecasting based elastic resource management in edge cloud," Computers & Industrial Engineering, vol. 139, pp. 1-16, 2020.
15. Jitendra Kumar, Ashutosh Kumar Singh, Rajkumar Buyya, "Self-directed learning-based workload forecasting model for cloud resource management," Information Sciences, vol. 543, pp. 345-366, 2021.
16. Durga S, Esther Daniel, Deepakanmani S, Reshma V.K, "Deep learning-based workload prediction and resource provisioning for mobile edge-cloud computing in healthcare applications," Sustainable Computing: Informatics and Systems, vol. 47, 2025.
17. Oumaima Ghandour, Said El Kafhali, Mohamed Hanini, "Adaptive workload management in cloud computing for service level agreements compliance and resource optimization," Computers and Electrical Engineering, vol. 120, 2024.
18. Yashwant Singh Patel, Jatin Bedi, "MAG-D: A multivariate attention network-based approach for cloud workload forecasting," Future Generation Computer Systems, vol.142, pp. 376-392, 2023.
19. Siyuan Liu, Qian He, Yiting Chen, Fan Zhang, "Wavelet and VMD enhanced traffic forecasting and scheduling method for edge cloud networks," Computers and Electrical Engineering, vol. 121, pp. 1-18, 2025.
20. Tahseen Khan, Wenhong Tian, Shashikant Ilager, Rajkumar Buyya, "Workload forecasting and energy state estimation in cloud data centres: ML-centric approach," Future Generation Computer Systems, vol. 128, pp. 320-332, 2022.
21. Sima Jeddi, Saeed Sharifian, "A hybrid wavelet decomposer and GMDH-ELM ensemble model for Network function virtualization workload forecasting in cloud computing," Applied Soft Computing, vol. 88, pp.1-16, 2020.
22. Weiping Zheng; Zongxiao Chen; Kaiyuan Zheng; Weijian Zheng; Yiqi Chen; Xiaomao Fan, "WorkloadDiff: Conditional Denoising Diffusion Probabilistic Models for Cloud Workload Prediction," IEEE Transactions on Cloud Computing, vol. 12, no. 4, pp. 1291-1304, 2024.
23. https://data.mendeley.com/datasets/b7bp6xhrcd/1