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

A Hybrid CatBoost–ExtraTrees Framework for Workload Variability Classification and QoS Prediction in Multi-Cloud Financial Environments

K. Bhaskar Babu[1], Thoutam Roshini[2], Vankudoth Eeswar[2], Nukaraju Sai Krishna Prudhvi Raj[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

ABSTRACT

Modern financial systems increasingly rely on multi-cloud environments to manage diverse and dynamic workloads with enhanced scalability, flexibility, and cost efficiency. However, maintaining consistent Quality of Service (QoS) across distributed platforms remains challenging due to fluctuating resource demands, varying service conditions, and strict Service Level Agreement (SLA) constraints. Efficient workload management therefore requires intelligent, adaptive mechanisms that ensure optimal resource utilization while avoiding over- or under-provisioning. To address these challenges, this study proposes a hybrid ensemble framework combining Categorical Boost (CB) and Extra Trees (ET) for workload variability classification and QoS score prediction. The framework adopts a two-stage learning approach integrating regression and classification tasks. In the first stage, regression models predict QoS scores using critical system performance metrics such as CPU utilization, memory usage, latency, throughput, and network bandwidth. In the second stage, the predicted QoS score is incorporated as an additional feature to classify workload variability into Low, Medium, and High categories. The system evaluates multiple machine learning models, including Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, alongside the proposed hybrid CB–ET ensemble. The implementation covers data preprocessing, exploratory data analysis, model training, performance evaluation, and visualization through an interactive Flask-based web interface. Experimental results show that while traditional models achieve moderate performance, the hybrid ensemble significantly improves both regression precision and classification accuracy. By leveraging CB's capability to model complex feature interactions and ET's randomness, the framework enhances generalization, robustness, and prediction stability, providing an effective solution for intelligent workload management and QoS optimization in multi-cloud financial environments.

Keywords: Multi-cloud computing, Quality of Service (QoS), Service Level Agreement (SLA), Hybrid ensemble learning, CatBoost, Extra Trees, Workload variability classification, QoS prediction, Resource optimization, Machine learning.

1. INTRODUCTION

Modern financial institutions increasingly relied on distributed multi-cloud computing infrastructures to support large-scale digital banking, payment processing, fraud detection, and real-time transaction processing systems. These environments enabled deployment across multiple cloud providers, improving scalability, reliability, and cost efficiency. However, the rapid growth in users, transactions, and services introduced significant workload variability, resulting in dynamic fluctuations in key performance metrics such as CPU utilization, memory usage, response time, service latency, and throughput. Such variability directly impacted the Quality of Service (QoS) delivered by financial applications, making consistent performance management challenging. Additionally, unpredictable workload patterns often led to inefficient resource allocation, causing over-provisioning or under-utilization of cloud resources. Therefore, effective monitoring, intelligent workload classification, and

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

accurate QoS prediction became essential for ensuring stable, efficient, and reliable service delivery in multi-cloud financial environments.

Historically, cloud infrastructure management relied on conventional techniques such as rule-based threshold monitoring, static workload allocation, and manual supervision. In these methods, administrators predefined limits for parameters like CPU utilization, memory usage, network bandwidth, and storage capacity, triggering scaling or redistribution actions when thresholds were exceeded. While such approaches provided basic operational control in earlier systems, they proved inadequate for modern multi-cloud financial environments characterized by highly dynamic and unpredictable workloads. The rapid variation in user demand and complex inter-service dependencies made static rules ineffective in capturing subtle performance changes. As a result, these methods often led to inefficient resource utilization, delayed response times, and inconsistent Quality of Service (QoS), highlighting the need for more adaptive and intelligent workload management strategies. As shown in figure 1 multi-cloud payment platform architecture for financial institutions illustrates deployment across Amazon Web Services (Amazon EKS), Microsoft Azure (Azure AKS), and Google Cloud (Google GKE). The architecture integrates API gateway (Kong), messaging system (NATS JetStream), and distributed database (CockroachDB**) to support highly available payment services and connectivity with external payment infrastructures.

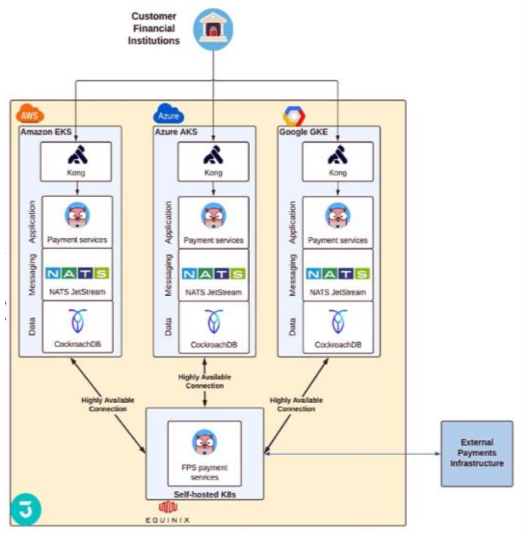

Figure 1: Multi-cloud payment resource platform

In financial systems where transaction processing, real-time analytics, and customer services operate simultaneously, maintaining stable service performance across multiple cloud providers is essential. Variations in workload intensity may occur due to peak transaction hours, sudden increases in online activity, or large-scale financial operations. Monitoring such variations and understanding their impact on service performance helps organizations ensure reliability, maintain regulatory compliance, and deliver consistent digital services to users.

2. LITERATURE SURVEY

Shabina Ghafir et al. [1] introduced an Intelligent particle swarm optimization (PSO)-based Feedback Controller designed for load balancing in cloud computing. The method focuses on enhancing task scheduling, resource allocation, and virtual machines (VMs) migration efficiency. Salimian et al. [2] proposed an evolutionary multiobjective optimization framework to address the IoT service placement problem in fog-enabled networks. Their method is based on a fog-cloud control middle ware and incorporates the MADE-k model (Monitoring, Analysis, Decision-making, and Execution) to optimize resources. They employed PSO to autonomously manage and deploy services, focusing on maximizing fog resource utilization and improving QoS by reducing Response Time and Service Cost.

Shahidinejad et al. [3] proposed a context-aware, multi-user offloading method for Mobile Edge Computing (MEC) that incorporates Federated Learning (FL) and Deep Reinforcement Learning (DRL). The suggested approach employs the MAPE loop (Monitor, Analyze, Plan, Execute) for autonomous resource management and integrates FL to enable collaborative learning among mobile devices and edge servers. Garima Verma et al. [4] proposes an optimization algorithm to improve resource allocation and load balancing in cloud environments. This study highlights the challenges of scheduling tasks across VMs in real-time dynamic settings.

Fatemeh Jazayeri et al. [5] proposed a Hidden Markov Model Auto-scaling Offloading (HMAO) method to optimize computation offloading in mobile fog computing (MFC). The method employs a Noninternalized-Plan-Execute (MAPE) control loop for resource management and utilizes a Hidden Markov Model (HMM) to predict module arrival rates and determine the optimal destination for offloading-whether to a Fog device or the Cloud. Seyed Salar Sefati et al. [6] proposed a novel approach to load balancing in cloud computing using the Grey Wolf Optimization (GWO) algorithm, focusing on maintaining resource reliability and enhancing QoS.

Singhal et al. [7] proposed a Rock Hyrax-based load balancing algorithm for cloud computing to address uneven load distribution, energy inefficiency, and local maxima issues. This method employs a probabilistic load estimation and fitness-based decisionmaking inspired by the foraging behavior of Rock Hyraxes, dynamically allocating tasks to VMs based on QoS parameters such as makespan, energy efficiency, throughput, and response time. Elsakaan et al. [8] proposed a novel multi-level hybrid algorithm for load balancing and task scheduling in cloud computing. Mosayebi et al. [9] introduce cost-efficient load balancing algorithms for cloud computing using the Artificial Immune System (AIS) and Clonal Selection Algorithm (CSA). These methods dynamically distribute workloads among VMs to optimize resource utilization, minimize response times, and reduce costs.

Ayothi et al. [10] proposed a hybrid dingo and whale optimization algorithm (HDWOA-LBM) for load balancing in cloud computing environments. The method combines the hunting characteristics of dingo optimization (DOA) with the exploitation phase improvements from whale optimization (IWOA) to achieve balanced exploration and exploitation during task scheduling.Alhijawi et al. [11] presented a novel Genetic-Based Recommender System based on historical rating data and semantic information. It is important to note that the study makes a distinction between evaluating and developing a list of probable suggestions. The Best List of Items Using Genetic Algorithms (BLIGA) uses a genetic algorithm to choose the best selections for the current user. Therefore, everyone represents a list of candidates that have been recommended. Three fitness functions are used in the genetic-based recommendation system to assess people hierarchically. Semantic similarity between items is assessed by the first function, which considers semantic information about the objects in question. Using the second function, it can compare the experience of a customer’s satisfaction. The third function picks the best possible list of suggestions for the user based on the predicted ratings.

Chenyang Li et al. [12] presented QoS-aware Web Service Composition (QWSC) as a trendy subject in both industry and academia due to the recent surge in the number of web facilities. QWSC has employed a meta-heuristic approach to solving classical optimization problems. Due to its inherent flaws, such a method generally fails in large-scale situations. For example, certain meta-heuristic

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

algorithms work well in nonstop search space, whereas QWSC's search space is discrete. Fang Li et al. [13] discussed meeting user needs for network worth of benefit; this research presents a novel QoS assessment prototype established on the Intelligent Water Droplets (IWD) algorithm. The network QoS assessment indices generated by the model are utilized to construct a multi-objective optimization function for addressing the issue. It is then utilized to investigate the model's most critical components using mathematical simulation. The simulation results suggest that the technique is more flexible than previous approaches.

Jiang et al. [14] suggested the use of cloud service tags and informative language to extract and identify latent connections between cloud services. Alayed et al. [15], in order to compose a web service, it is necessary to understand the demands of the end-user. Each task in this pipeline provides an abstract description of specific user demands. For workflow management, web services can be gathered from several sources. The candidate list refers to a collection of services provided by various service providers for a certain activity.

3. PROPOSED SYSTEM

The proposed system introduced an intelligent analytical framework for evaluating service performance and workload variability in multi-cloud financial environments. It systematically analysed key resource utilization parameters, including CPU usage, memory consumption, storage usage, network bandwidth, service latency, response time, throughput, and load balancing levels to estimate QoS and identify workload patterns. By integrating advanced analytical models into the cloud monitoring process, the system provided a structured and data-driven approach to assess service behaviour across multiple cloud platforms. This enabled more accurate performance evaluation and improved understanding of dynamic workload conditions. As illustrated in Figure 2, the approach enhanced resource planning, optimized service management, and improved the overall reliability and efficiency of financial applications operating in multi-cloud environments.

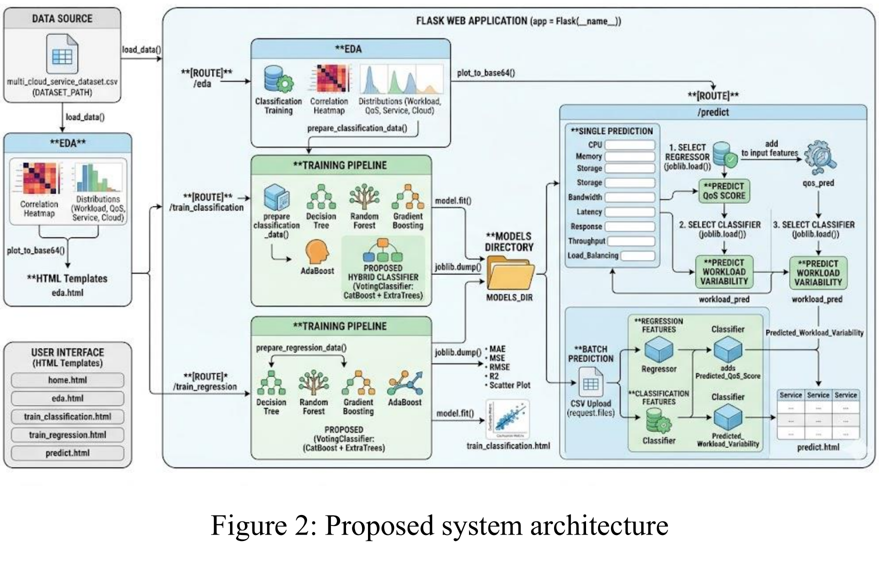

Figure 2: Proposed system architecture

1. Data Collection from Multi-Cloud Environment: The system collects operational data from different cloud service platforms, including metrics such as CPU utilization, memory usage, storage usage, network bandwidth, service latency, response time, throughput, and load balancing information. These parameters represent the performance characteristics of financial service workloads operating across distributed cloud infrastructures.

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

2. Data Organization and Preparation: The collected cloud performance data is organized into a structured dataset format where each record represents a service instance and its associated resource usage metrics. This structured representation allows the system to analyse system performance indicators and prepare the data for further evaluation.

3. QoS Score Estimation: Using the collected resource utilization parameters, the system evaluates and estimates the QoS score for each cloud service instance. The QoS score represents the overall service performance level based on operational metrics such as latency, response time, and resource utilization efficiency.

4. Workload Variability Evaluation: After estimating the QoS score, the system analyzes workload characteristics and categorizes the workload variability into predefined levels such as Low, Medium, or High. This classification helps understand how service demand changes across different cloud platforms.

5. Performance Monitoring and Result Visualization: Finally, the system presents the predicted QoS scores and workload variability levels through a monitoring interface. The results help administrators understand system behaviour, evaluate service performance trends, and support decision-making for efficient multi-cloud resource management.

4. RESULTS ANALYSIS

The results demonstrate the effectiveness of the proposed hybrid CB–ET ensemble framework in accurately predicting QoS scores and classifying workload variability in multi-cloud environments. Compared to traditional machine learning models, the hybrid approach achieves higher regression precision and improved classification accuracy across all workload categories. The integration of predicted QoS scores as an additional feature significantly enhances classification performance. Experimental analysis also shows better generalization and stability under varying workload conditions. Visualization through the Flask-based interface further validates model consistency and interpretability. The framework proves to be a robust and efficient solution for intelligent workload management and QoS optimization.

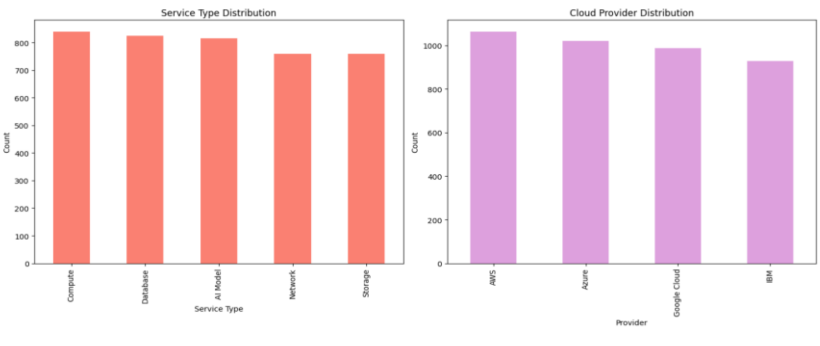

Figure 3: Distribution plots for Service type and Cloud provider columns

The figure 3 shows that the first graph depicts the Service Type Distribution, where different service categories such as Compute, Database, AI Model, Network, and Storage are fairly evenly distributed, indicating that the dataset includes diverse workload types without significant imbalance. The right graph represents the Cloud Provider Distribution, where providers like AWS, Azure, Google Cloud, and IBM are also distributed relatively uniformly, with AWS having slightly higher representation. This balanced distribution across service types and cloud providers ensures that the CART-based RT and

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

CA models can generalize well across different cloud environments and workload scenarios without bias toward any specific category.

Figure 4 illustrates the confusion matrix for the proposed CBET model shows excellent classification performance with near-perfect accuracy across all classes. It correctly predicts all Low (253) and High (279) instances without any misclassification, and almost all medium instances (260) with only a very small number misclassified as High (8). There are no cross-class errors between Low and other classes, indicating strong class separation. This demonstrates that the CBET model effectively captures complex patterns and significantly outperforms other models in distinguishing workload variability levels.

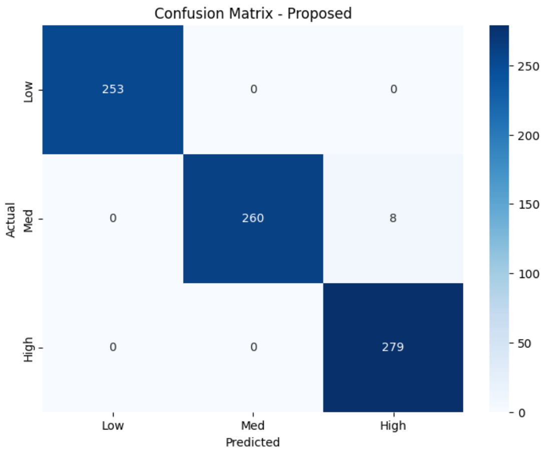

Figure 4: Illustration of confusion matrix using proposed CBET model for workload_variability class

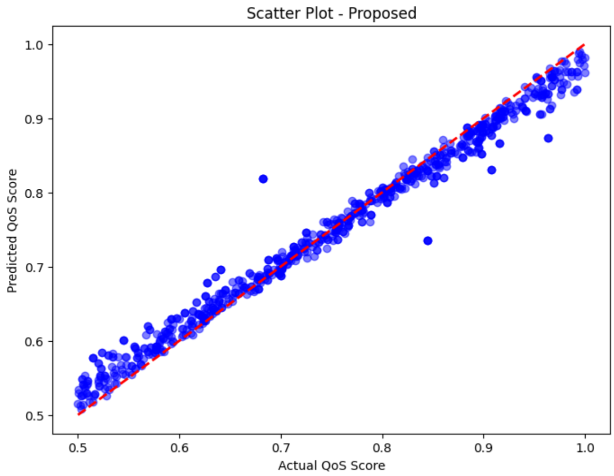

Figure 5: Scatter plot obtained using CBET model for QoS_score

Figure 5 shows scatter plot for the proposed CBET model shows a near-perfect alignment of predicted QoS values along the ideal diagonal line, indicating highly accurate predictions. The data points are tightly clustered around the line across the entire range, demonstrating that the model effectively captures both low and high QoS variations. Only a few minor deviations are observed, suggesting minimal prediction error. This reflects that the CBET model achieves superior regression performance with excellent generalization and precise modeling of the relationship between input features and QoS score.

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

Network_Bandwidth (Mbps)	Service_Latency (ms)	Response_Time (ms)	Throughput (Requests/sec)	Load_Balancing (%)	Predicted_QoS_Score	Predicted_Workload_Variability
650.25	175.50	400.75	850.25	80.50	0.815011	High
520.75	190.25	310.50	550.75	75.25	0.810352	Low
780.50	145.75	280.25	920.50	88.75	0.781269	Med

Figure 6: Prediction obtained on sample test data using proposed CBET Model

The figure 6 shows prediction results obtained from the proposed CBET model demonstrate effective performance in both QoS regression and workload classification. For the given input parameters, the model predicts QoS scores in the range of approximately 0.78 to 0.81, indicating stable and high service quality across different conditions. Based on these QoS values and input features, the model accurately classifies workload variability into different categories such as High, Low, and Medium. The variation in predictions shows that the model is sensitive to changes in network bandwidth, latency, response time, throughput, and load balancing. These results confirm that the CBET model successfully captures the relationship between cloud resource parameters and system performance, providing reliable and consistent predictions.

4.1 Comparative analysis

Table 1: Performance Comparison of Classification models for workload_variability class

Model	Accuracy	Precision	Recall	F-Score
Existing DTC	34.50%	42.96%	34.50%	18.88%
Existing RFC	41.75%	45.23%	41.75%	37.46%
Existing GBC	76.88%	77.23%	76.88%	76.87%
Existing AdaBoost	40.62%	40.15%	40.62%	36.86%
Proposed CBET	99.10%	99.03%	99.00%	99.03%

The table 1 shows the performance comparison of classification models for workload variability clearly highlights the superiority of the proposed CBET model over existing approaches. The existing DTC shows very poor performance with an accuracy of 34.50% and a low F-score of 18.88%, indicating weak class separation and high misclassification. The existing RFC slightly improves performance with 41.75% accuracy and a better F-score of 37.46%, but still struggles to generalize effectively. The Existing AdaBoost model performs similarly with 40.62% accuracy, showing limited improvement in classification capability. In contrast, the Existing GBC model demonstrates a significant improvement with 76.88% accuracy and balanced precision, recall, and F-score around 77%, indicating strong learning of class patterns. However, the proposed CBET model achieves outstanding performance with 99.10% accuracy, 99.03% precision, 99.00% recall, and 99.03% F-score, reflecting near-perfect classification with minimal errors. This dramatic improvement confirms that the hybrid CBET approach effectively captures complex workload patterns and significantly enhances classification accuracy compared to traditional CART-based models.

Table 2: Performance Comparison of Classification models for QoS_score

Model	MAE	MSE	RMSE	R2-Score

Existing DTC	0.1206	0.019	0.1401	0.0196
Existing RFC	0.1183	0.0189	0.1376	0.0542
Existing GBC	0.099	0.0141	0.1188	0.2952
Existing AdaBoost	0.1193	0.0192	0.1384	0.0427
Proposed CBET	0.0171	0.0006	0.0248	0.9692

The table 2 presents performance comparison of regression models for QoS score prediction clearly demonstrates the effectiveness of the proposed CBET model over existing methods. The existing DTC shows weak performance with higher error values (MAE: 0.1206, RMSE: 0.1401) and a very low R^2 score of 0.0196, indicating poor predictive capability. The existing RFC and AdaBoost models show slight improvements but still maintain high error rates and low R^2 scores (0.0542 and 0.0427 respectively), reflecting limited ability to capture the relationship between features and QoS. The Existing GBC performs comparatively better with reduced errors (MAE: 0.099, RMSE: 0.1188) and a moderate R^2 score of 0.2952, indicating improved learning of patterns. However, the proposed CBET model significantly outperforms all others with extremely low error values (MAE: 0.0171, RMSE: 0.0248) and a very high R^2 score of 0.9692, showing near-perfect prediction accuracy. This substantial improvement confirms that the CBET model effectively models complex relationships in the data and provides highly accurate QoS predictions compared to traditional CART-based approaches.

5. CONCLUSION

The research successfully presents an effective CART-based framework for analysing multi-cloud workload variability and predicting QoS scores using both classification and regression approaches. By integrating multiple models such as DTC, RFC, GBC, AdaBoost, and the proposed CBET, the system establishes a dual-stage pipeline where RT predicts QoS and CA utilizes it for workload classification. Experimental results demonstrate that traditional models like DTC, RFC, and AdaBoost show limited performance, while GBC provides moderate improvements in both tasks. However, the proposed CBET model significantly outperforms all existing models by achieving near-perfect classification accuracy and highly precise regression predictions with minimal error. The system also ensures robust implementation through proper preprocessing, EDA, visualization, and a Flask-based interface for real-time predictions. The research proves that combining CB and ET in a hybrid CART framework enhances model generalization, improves prediction reliability, and provides an efficient solution for intelligent multi-cloud workload management and performance optimization.

REFERENCES

[1]. Ghafir, S., Alam, M.A., Siddiqui, F., Naaz, S.: Load balancing in cloud computing via intelligent pso-based feedback controller. Sustain. Comput. Inform. Syst. 41, 100948 (2024)

[2]. Salimian, M., Ghobaei-Arani, M., Shahidinejad, A.: An evolutionary multi-objective optimization technique to deploy the iot services in fog-enabled networks: an autonomous approach. Appl. Artif. Intell. 36(1), 2008149 (2022)

[3]. Shahidinejad, A., Farahbakhsh, F., Ghobaei-Arani, M., Malik, M.H., Anwar, T.: Context-aware multi-user offloading in mobile edge computing: A federated learning-based approach. J. Grid Comput. 19(2), 18 (2021).

[4]. Verma, G.: Load balancing in cloud environment using opposition based spider monkey optimization. Wirel. Pers. Commun. 137(2), 977–996 (2024).

[5]. Jazayeri, F., Shahidinejad, A., Ghobaei-Arani, M.: A latency-aware and energy-efficient computation offloading in mobile fog computing: a hidden markov model-based approach. J. Supercomput. 77, 4887–4916 (2021)

[6]. Sefati, S., Mousavinasab, M., Zareh Farkhady, R.: Load balancing in cloud computing environment using the grey wolf optimization algorithm based on the reliability: performance evaluation. J. Supercomput. 78(1), 18–42 (2022)

[7]. Singhal, S., Sharma, A., Verma, P.K., Kumar, M., Verma, S., Kaur, M., Rodrigues, J.J., Khurma, R.A., García-Arenas, M., et al.: Energy efficient load balancing algorithm for cloud computing using rock hyrax optimization. IEEE Access (2024)

[8]. Elsakaan, N., Amroun, K.: A novel multi-level hybrid load balancing and tasks scheduling algorithm for cloud computing environment. J. Supercomput. 1–41 (2024)

[9]. Mosayebi, M., Azmi, R.: Cost-effective clonal selection and ais-based load balancing in cloud computing environment. J. Supercomput. 80(16), 23271–23310 (2024).

[10]. Ramya, K., Ayothi, S.: Hybrid dingo and whale optimization algorithm-based optimal load balancing for cloud computing environment. Trans. Emerg. Telecommun. Technol. 34(5), e4760 (2023).

[11]. Alhijawi, B.; Kilani, Y. A collaborative filtering recommender system using genetic algorithm. Inf. Process. Manag. 2020, 57, 102310.

[12]. Li, C.; Li, J.; Chen, H. A Meta-Heuristic-Based Approach for Qos-Aware Service Composition. IEEE Access 2020, 8, 69579–69592.

[13]. Li, F.; Wang, Y.; Li, P. The Evaluation Model of Network QoS Based on Intelligent Water Droplets Algorithm. Discret. Dyn. Nat. Soc. 2020, 2020, 1–7.

[14]. Jiang, Y.; Tao, D.; Liu, Y.; Sun, J.; Ling, H. Cloud service recommendation based on unstructured textual information. Futur. Gener. Comput. Syst. 2019, 97, 387–396.

[15]. Alayed, H.; Dahan, F.; Alfakih, T.; Mathkour, H.; Arafah, M. Enhancement of Ant Colony Optimization for QoS-Aware Web Service Selection. IEEE Access 2019, 7, 97041–97051.