An Attention-Based LSTM-ELECTRE Model for Intelligent and Proactive Load Balancing in Real-Time Fog Computing Environments

Abrar S. Kadhim
College of Information Technology, University of Babylon, Iraq
inf365.abrar.saad@uobabylon.edu.iq (corresponding author)

Received: 26 May 2025 | Revised: 6 July 2025 | Accepted: 27 July 2025

 DOI: https://doi.org/10.48084/etasr.12374

ABSTRACT

Load balancing between nodes is a fundamental challenge for fog computing environments, which are associated with dynamic Internet of Things (IoT) applications that require real-time responses. From this perspective, this study proposes an advanced hybrid model that combines the deep learning capabilities of LSTM networks with an attention mechanism to focus on the most significant time segments of the data series when predicting the upcoming load. Then, a multi-criteria decision-making algorithm follows, the ELECTRE algorithm, allowing optimal load prediction and intelligent task distribution across nodes to achieve fog computing stability. The proposed model bridges the gap in previous studies that do not consider task distribution mechanisms across nodes, load balancing prediction, and the nonlinear nature of the data. The proposed model was tested on two Mendeley datasets and a synthetic dataset to evaluate its generalization performance. The proposed model demonstrated excellent performance in terms of accuracy (R^2 up to 0.99), low prediction error (MAE = 0.0085), effective load balance, and 100% task completion rate. The results obtained, compared to other approaches, confirm that the proposed method demonstrates high accuracy, is generalizable, and can be integrated with other models, such as fault prediction systems. This makes it a promising framework for more stable and reliable fuzzy environments within the current timeframe.

Keywords-fog computing; proactive load balancing; LSTM; attention mechanism; ELECTRE; multi-criteria decision analysis; real-time systems; resource optimization; Internet of Things (IoT); Quality of Service (QoS)

I. INTRODUCTION

In recent years, the world has witnessed an explosion in the number of Internet of Things (IoT) devices, leading to a growing need for distributed computing architectures that are more responsive to time-sensitive environments. In this context, fog computing has emerged as an intermediate layer between cloud data centers and edge devices, allowing computation, storage, and network services to be performed locally near data sources [1]. This architecture contributes to reduced response time and reduced bandwidth pressure, making it suitable for real-time applications such as smart healthcare, factory automation, and connected vehicles [2, 3].

With the large distribution of nodes in fog computing environments, load balancing becomes a key challenge to ensure optimal resource utilization and maintain Quality of Service (QoS). Many traditional approaches, such as Round Robin and nearest node selection, rely on instantaneous decisions that lack awareness of the future state of the system, limiting their ability to adapt to dynamic conditions and often leading to bottlenecks at some nodes [4]. In light of these shortcomings, recent developments in artificial intelligence have demonstrated the importance of adopting machine learning and deep learning models to provide intelligent, predictive solutions. Long-Short-Term Memory (LSTM) networks have emerged as effective tools for analyzing temporal data and predicting future load trends and response delays, making them ideal for environments that require precise, advanced decisions [5-7].

This paper proposes an advanced predictive load balancing model that combines LSTM with an attention mechanism and the ELECTRE multi-criteria decision-making algorithm to enhance prediction accuracy by focusing the model on the most significant patterns in load data. This study employs a real-world dataset of load distributions [8], a three-layer experimental data from a smart home environment [9], and an artificially generated dataset to evaluate the comprehensiveness and efficiency of the model. The proposed model distinguishes itself from other traditional or partial models by its integration of prediction and decision-making, giving it the ability to reduce delays, improve load balance, and increase the rate of successful task completion in dynamic and realistic fog computing environments.

In [10], a new method was proposed for a real-world dataset of 10 web services. This method used ARIMA and ARFIMA to analyze long-memory time series to predict cloud service performance. The results showed that ARFIMA

outperformed ARIMA in the prediction of response time, reducing the prediction error by up to 57.8%. However, these models did not address load balancing or multi-criteria decision-making and were focused solely on cloud computing environments. In [8], fuzzy logic was used to distribute load among nodes in a fog computing environment. This study presented a lightweight and fast algorithm for load balancing in a fog computing environment, but it could not predict load in advance, leading to less accurate node selection. In [11], ARIMA was applied to predict future loads in a cloud computing environment, and results showed that it can predict future loads, which helps improve the Quality of Service (QoS). In [12], AHP and TOPSIS were used on various real datasets to evaluate and improve load balancing in fog computing networks. The results demonstrated high accuracy in selecting nodes at the current moment, while they did not provide load prediction and were not time-adaptive.

In [13], a deep neural network with GRU was used to predict loads in a real-world cloud computing environment (Alibaba and Google). The results demonstrated an accurate prediction of multivariate load, with a 15% reduction in mean square error compared to a GRU-based approach alone. Multiple studies explored similar directions in predictive load balancing and decision-making models. For instance, in [14], an artificial neural network (ESANN) was used to predict loads and improve SLA performance, demonstrating accurate load prediction and improving SLA performance in a cloud computing environment. Similarly, in [15], LSTM was combined with CRP for proactive fault prediction in fog computing devices. This approach was evaluated on experimental data, achieving a prediction accuracy of 98.69% while reducing processing time and improving fault prediction accuracy without requiring decision-making or load balancing. In [16], a multilevel ensemble model used multiple machine learning techniques to classify node status into three categories, overloaded, balanced, and underloaded, providing an accurate prediction of node status in a fog computing environment. In [17], Bayesian deep learning models were used to forecast future loads while estimating uncertainty. The results demonstrated that prediction models that take uncertainty into account lead to improved performance, especially in service-level indicators. In [18], multi-criteria decision analysis techniques such as FAHP and FTOPSIS were used to select optimal nodes in fog and cloud computing environments. The results showed improved load balancing, reduced response time, and increased resource utilization by up to 90%. However, this study relied solely on multi-criteria decision analysis without incorporating a predictive component.

II. PROPOSED METHODOLOGY

To achieve the goal of improving the load distribution in fog computing environments by predicting loads before they occur, the proposed model relies on a hybrid architecture that combines prediction (through an attention-assisted LSTM algorithm) and robust decision-making (ELECTRE). Figure 1 illustrates the flowchart of the proposed method, which is implemented through the following steps.

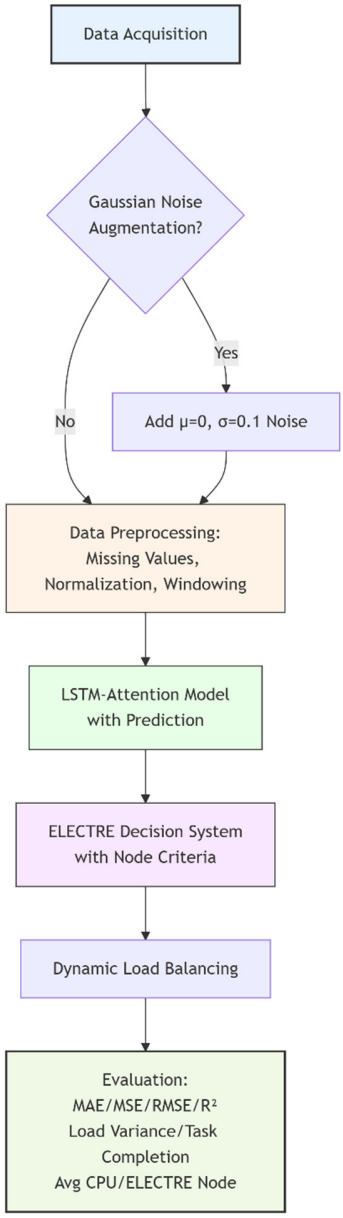

Fig. 1. Flowchart for the proposed method.

A. Data Acquisition

The model begins with data preparation. Three types of datasets were used. Two datasets represented different scenarios in fog computing environments:

- The Distributions Dataset for Fuzzy-Based Fog Load Balancer [19].
- The Experimental datasets of three-tier fog computing-based gateway nodes for smart home automation applications [20].

The third dataset was artificially generated to cover additional scenarios, ensuring the model's comprehensiveness across diverse situations and enhancing its generalization ability.

B. Data Preprocessing

The data were prepared before running the model, addressing problems that may arise during algorithm execution or prediction. This proposed method performs preprocessing operations, including handling missing values using time-interpolation algorithms, data normalization using a MinMax Scaler within the domain, and windowing, where each input sample is constructed from $t-n$ and prediction is performed in the $t+1$ step. In cases where the data is limited in samples, the Gaussian noise principle is used to generate additional data and increase the dataset size to improve model performance during training without compromising the integrity of the original data, and only when the data is limited.

C. Model Design and Prediction

The proposed model consists of two LSTM layers, with long-term memory followed by an attention layer to identify periods with the highest impact on the prediction. The average of the context vectors is used as input to a dense layer, which produces a predicted value for the future load. After determining the predicted values, they are sent to the ELECTRE system, which evaluates each node based on the expected load, expected response time, energy consumption, delay sensitivity, and location or geographical distance (if available). Thus, the weights are formed, which are transformed into a concordance matrix according to:

$$C(i,j) = \sum_{\{k=1\}}^{n} w_k \cdot \delta\left(x_{\{ik\}} \geq x_{\{jk\}}\right) \quad (1)$$

where $C(i,j)$ represents the concordance index, n represents the number of criteria used in the evaluation, k is an indicator that represents each of the criteria ($k = 1, 2, \dots, n$), w_k represents the relative weight of the criterion k, x_{ik} is the normalized value obtained by the alternative i at the criterion k, x_{jk} is the value obtained by alternative j at criterion k, and $\left(x_{\{ik\}} \geq x_{\{jk\}}\right)$ is the indicator function, which takes the value 1 or 0. Thus, the optimal node with the highest compatibility score compared to the rest of the nodes is selected.

D. Model Evaluation

Evaluation relied on the most important and rigorous criteria to measure the accuracy and quality of the prediction metrics to ensure that the model performs efficiently, as well as the possibility of comparing it with previous works and models to measure the rate of improvement. Evaluation relied on a set of indicators, including:

- MAE, MSE, RMSE, and R^2 scores to assess the accuracy of load prediction.
- Load variance is an indicator of distribution balance.
- Average response time is a measure of service quality.
- Task completion rate to determine scheduling efficiency.
- Average energy consumption is a sustainability indicator when energy data is available.

III. PERFORMANCE EVALUATION

A. Evaluation Metrics and Experimental Findings

The proposed hybrid model was tested on two public and a generated dataset. One was a real-world dataset for load distribution in a general fog environment, while the second simulated a complex, three-layered Edge-Fog-Cloud system for smart home applications [19, 20]. The third dataset was generated artificially to represent a scenario designed to test the model's generalization efficiency across different environments.

Several important metrics were measured to evaluate the model's efficiency, including mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), coefficient of determination (R^2), load variance, task completion rate (TCR), and average power or processor consumption. The optimal node for task execution was identified using the multi-criteria ELECTRE method. Table I represents the results of the proposed model for each dataset.

TABLE I. RESULTS OF THE PROPOSED MODEL FOR EACH DATASET

Dataset	MAE	MSE	RMSE	R^2	Var.	Task %	Avg.	BN
[19]	0.0500	0.0050	0.0700	0.9665	0.0100	100%	0.40	N1
[20]	0.0140	0.0003	0.0183	0.9781	0.0146	100%	27.5804	N0
Synthetic	0.0085	0.00012	0.0109	0.9938	0.0152	100%	5.8013	N5

The results shown in Table I indicate the superiority of the proposed model in terms of accuracy and reliability.

- The MAE, MSE, and RMSE values were very low, indicating the quality of the predictions in different environments.
- The coefficient of determination (R^2) maintained its very high values, ranging between 0.9665 and 0.9938, indicating the model's ability to represent relationships excellently.
- The model maintained 100% task completion in all scenarios, demonstrating the effectiveness of the ELECTRE algorithm in selecting the optimal node based on multi-criteria measurements.
- The model showed a low load variance, indicating a fair and balanced distribution of tasks among the nodes.
- Despite the variation in power/processor consumption between environments, operational efficiency was maintained without compromising performance.

B. Analytical Assessment of the Results

All results obtained from the three scenarios were transformed into graphical representations to evaluate the performance and behavior of the proposed model. Figure 2 shows the error indices (MAE, MSE, RMSE) in load prediction. All indices recorded a significant decline in both synthetic and real values, demonstrating the model's ability to generalize to different environments.

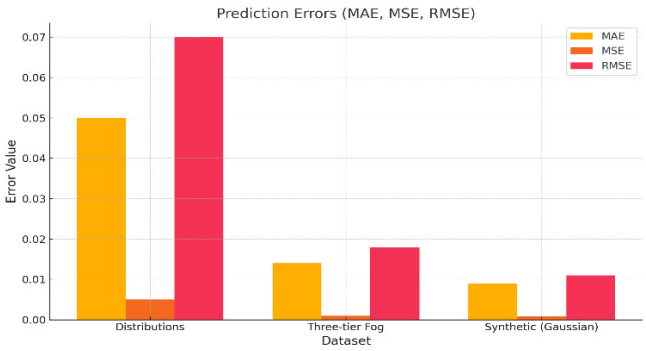

Fig. 2. Error indicators (MAE, MSE, RMSE) in load prediction.

Figure 3 shows the R^2 scores, with all values exceeding 96% and even reaching 99%. This demonstrates the high accuracy of prediction and stable performance regardless of the nature of the data used.

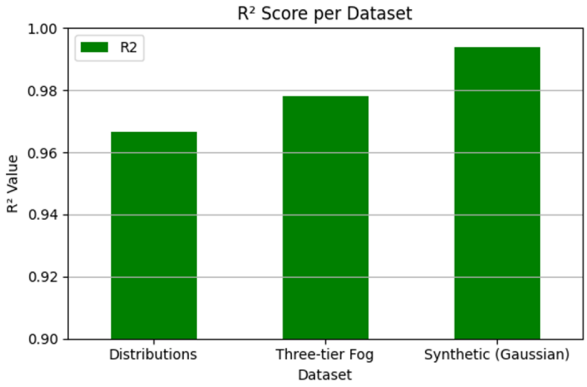

Fig. 3. R^2 score for each dataset.

Figure 4 illustrates the load variance, showing that the variance in load distribution is relatively low in all scenarios, indicating a balanced distribution of tasks among all nodes. This metric is one of the most important indicators, as it is linked to reducing latency and improving service quality.

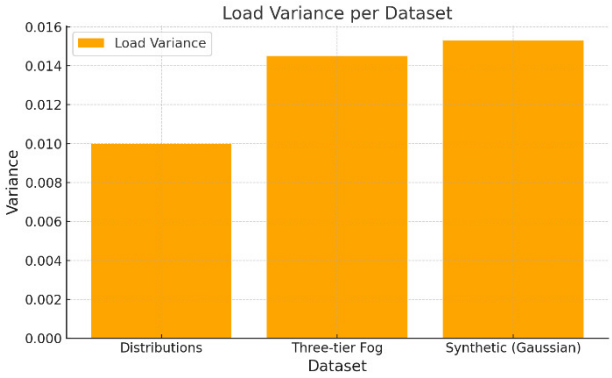

Fig. 4. Load variance per dataset.

Figure 5 shows that all scenarios completed tasks at a 100% rate, demonstrating excellent responsiveness, efficient prediction, and the ability to maintain excellent service quality at all times.

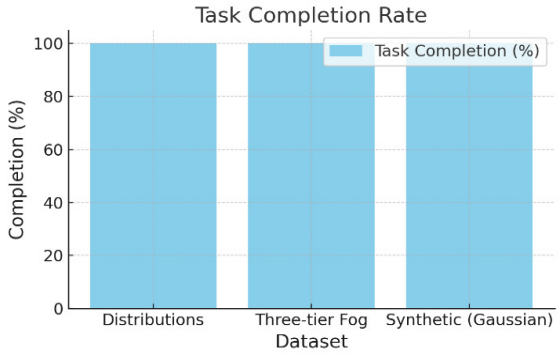

Fig. 5. Task completion rate across datasets.

Figure 6 shows that the average power consumption or CPU consumption decreases in the synthetic and distributed dataset scenarios, while it increases slightly in the smart home dataset scenario due to its large dataset and the computational intensity it requires. However, this does not affect the efficiency of the other evaluations.

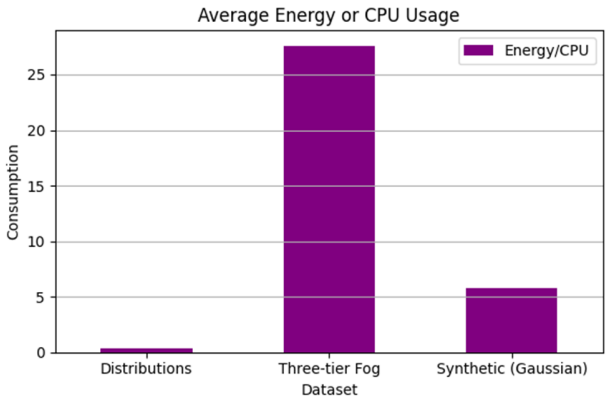

Fig. 6. Average power consumption of the CPU.

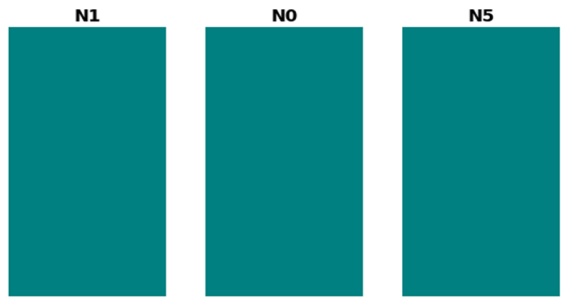

Fig. 7. Best node selected by the ELECTRE algorithm.

Figure 7 shows the ideal nodes selected by the ELECTRE algorithm, which vary across scenarios depending on their nature, data used, and load. This diversity reflects the model's

accuracy in evaluating variables based on environmental context, resource consumption, and response time.

C. Comparative Performance Analysis with Existing Studies

The proposed model demonstrated high predictive accuracy and remarkable stability in load distribution across nodes, making it a reliable and scalable model for fuzzy computing. A performance comparison with [8] on the same distribution dataset shows that the proposed method demonstrated significant improvements in accuracy. It also demonstrated superior distribution stability and a task completion rate of 100%, a level previously unachieved at this level. Performance comparison using [20] outperformed the approach presented in [12]. Although the analytical tools are similar, since this study also relies on ELECTRE, the lack of a predictive component limits its ability to accurately predict load fluctuations, making the proposed approach more effective and flexible when dealing with realistic foggy environments.

The proposed approach also complements the approach in [15], which explores fault prediction using LSTM within a policy-based framework. Combining the two models creates a comprehensive solution that combines fault prediction and proactive load balancing, mitigating faults before they occur and ensuring system stability under increasing demand and changing loads. The proposed model also outperformed the method in [11], which used the ARIMA model to provide an ensemble prediction model without incorporating a multi-criteria decision-making component, providing higher prediction accuracy. This is a key strength of the proposed approach, which combines prediction accuracy with decision wisdom. Table II shows a detailed comparison between the proposed model and previous studies in terms of the most important points and an indication of their strengths and weaknesses.

TABLE II. COMPARATIVE ANALYSIS OF RELATED WORKS

Ref.	Dataset	Method	Key Strengths	Weaknesses
Fuzzy Load Balancer [8]	Real (same)	Fuzzy Only	Fast, lightweight	No prediction, less accurate node selection
MCDA-based Resilience [12]	Real (other)	MCDA (AHP, TOPSIS)	Accurate short-term decisions	No forecasting, time-insensitive
LSTM fault prediction [15]	Experimental	LSTM	Good failure prediction	No decision-making or balancing
ARIMA QoS prediction [11]	Real (cloud)	ARIMA	Simple, efficient	Not suitable for non-linear loads
ML ensemble prediction [16]	Real	ML Ensemble	Accurate node load prediction	No integrated balancing logic
Proposed Model	Real+Experimental	Predict + Decide	High accuracy ($R^2 > 0.96$), adaptive, integrated	No current limitations in tests

IV. CONCLUSION

This study presented a hybrid model to predict and optimize load balancing in fog computing environments, combining the predictive strength of LSTM with an attention mechanism, which is important for identifying important time series, and the decision-making accuracy across various important parameters using the ELECTRE algorithm. Tested in various fog computing environments, the results showed that this combination of AI techniques and decision analysis methods enhances the efficiency of load balancing and contributes to reducing error rates and performance variability for nodes, even in complex and changing fog environments. The model outperformed previously used models not only in synthetic data, but also in real data collected from operational environments, demonstrating its practical applicability in IoT and real-time systems. It also demonstrated excellent integration with other predictive approaches, paving the way for the development of comprehensive proactive frameworks that combine fault prediction and load balancing. These results represent a significant step toward building smarter and more adaptive fog computing environments and lay the foundation for promising future research avenues, including improved response time, integration with reinforcement learning techniques, and privacy-enhancing decision-making The novelty of the proposed Attention-LSTM-ELECTRE model lies in its hybrid design that combines time-series forecasting with multi-criteria decision analysis, enabling proactive and intelligent task distribution in fog environments. This integration significantly improves responsiveness and resource optimization, making it highly applicable to latency-sensitive and resource-constrained IoT applications.

DATA AVAILABILITY

This study used two publicly available real-world datasets and one synthetic dataset generated for validation purposes. The first dataset, Distributions Dataset for Fuzzy Based Fog Load Balancer, is accessible through Mendeley Data [19]. The second dataset, Experimental datasets of three-tier fog computing-based gateway node for the smart home automation applications, is also available through Mendeley Data [20]. The synthetic dataset, created through controlled Gaussian noise expansion for robust testing, is not publicly available but can be provided upon reasonable request for academic and non-commercial use.

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, May 2012, pp. 13–16, https://doi.org/10.1145/2342509.2342513.

[2] S. Yi, C. Li, and Q. Li, "A Survey of Fog Computing: Concepts, Applications and Issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*, Mar. 2015, pp. 37–42, https://doi.org/10.1145/2757384.2757397.

[3] A. V. Dastjerdi and R. Buyya, "Fog Computing: Helping the Internet of Things Realize Its Potential," *Computer*, vol. 49, no. 8, pp. 112–116, Dec. 2016, https://doi.org/10.1109/MC.2016.245.

[4] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal Workload Allocation in Fog-Cloud Computing Toward Balanced Delay and Power Consumption," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1171–1181, Sep. 2016, https://doi.org/10.1109/JIOT.2016.2565516.

[5] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590–3605, Sep. 2016, https://doi.org/10.1109/JSAC.2016.2611964.

[6] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017, https://doi.org/10.1109/ACCESS.2017.2685434.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Aug. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[8] S. P. Singh, A. Sharma, and R. Kumar, "Design and exploration of load balancers for fog computing using fuzzy logic," *Simulation Modelling Practice and Theory*, vol. 101, May 2020, Art. no. 102017, https://doi.org/10.1016/j.simpat.2019.102017.

[9] M. M. Rathore, A. Ahmad, A. Paul, and U. K. Thikshaja, "Exploiting real-time big data to empower smart transportation using big graphs," in *2016 IEEE Region 10 Symposium (TENSYMP)*, Bali, Indonesia, May 2016, pp. 135–139, https://doi.org/10.1109/TENCONSpring.2016.7519392.

[10] H. Nourikhah, M. K. Akbari, and M. Kalantari, "Modeling and predicting measured response time of cloud-based web services using long-memory time series," *The Journal of Supercomputing*, vol. 71, no. 2, pp. 673–696, Feb. 2015, https://doi.org/10.1007/s11227-014-1317-4.

[11] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 449–458, Jul. 2015, https://doi.org/10.1109/TCC.2014.2350475.

[12] M. Ebrahim and A. Hafid, "Resilience and load balancing in Fog networks: A Multi-Criteria Decision Analysis approach," *Microprocessors and Microsystems*, vol. 101, Sep. 2023, Art. no. 104893, https://doi.org/10.1016/j.micpro.2023.104893.

[13] M. Xu, C. Song, H. Wu, S. S. Gill, K. Ye, and C. Xu, "esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 3, May 2022, https://doi.org/10.1145/3524114.

[14] A. Gupta and H. S. Bhadauria, "Workload prediction for SLA performance in cloud environment: ESANN approach," *Intelligent Decision Technologies*, vol. 17, no. 4, pp. 1085–1100, Nov. 2023, https://doi.org/10.3233/IDT-230101.

[15] H. Sabireen and N. Venkataraman, "Proactive Fault Prediction of Fog Devices Using LSTM-CRP Conceptual Framework for IoT Applications," *Sensors*, vol. 23, no. 6, Jan. 2023, Art. no. 2913, https://doi.org/10.3390/s23062913.

[16] S. Bawa, P. S. Rana, and R. Tekchandani, "Multilevel Ensemble Model for Load Prediction on Hosts in Fog Computing Environment," *Computing and Informatics*, vol. 43, no. 5, pp. 1053–1083, Oct. 2024, https://doi.org/10.31577/cai_2024_5_1053.

[17] A. Rossi, A. Visentin, D. Carraro, S. Prestwich, and K. N. Brown, "Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning," *Cluster Computing*, vol. 28, no. 4, Feb. 2025, Art. no. 258, https://doi.org/10.1007/s10586-024-04933-2.

[18] A. A. A. Gad-Elrab, A. S. Alsharkawy, M. E. Embabi, A. Sobhi, and F. A. Emara, "Adaptive multi-criteria-based load balancing technique for resource allocation in fog-cloud environments," *International journal of Computer Networks & Communications*, vol. 16, no. 1, pp. 105–124, Jan. 2024, https://doi.org/10.5121/ijcnc.2024.16107.

[19] S. P. Singh, "Distributions Dataset for Fuzzy Based Fog Load Balancer." Mendeley, Sep. 28, 2019, https://doi.org/10.17632/FZ8RSR7C2K.1.

[20] K. Habib, "Experimental datasets of three-tier fog computing based gateway node for the smart home automation applications." Mendeley, Sep. 06, 2022, https://doi.org/10.17632/8J3WNN5DCC.1.