Enhanced Virtual Machine Resource Optimization in Cloud Computing Using Real-Time Monitoring and Predictive Modeling

Rim Doukha, Abderrahmane Ez-zahout
Intelligent Processing and Security of Systems Team-Faculty of Sciences, Mohammed V University, Rabat, Morocco

***Abstract*—Effective resource estimation is essential in cloud computing to minimize operational costs, optimize performance, and enhance user satisfaction. This study proposes a comprehensive framework for virtual machine optimization in cloud environments, focusing on predictive resource management to improve resource efficiency and system performance. The framework integrates real-time monitoring, advanced resource management techniques, and machine learning-based predictions. A simulated environment is deployed using PROXMOX, with Prometheus for monitoring and Grafana for visualization and alerting. By leveraging machine learning models, including Random Forest Regression and LSTM, the framework predicts resource usage based on historical data, enabling precise and proactive resource allocation. Results indicate that the Random Forest model achieves superior accuracy with a MAPE of 2.65%, significantly outperforming LSTM's 17.43%. These findings underscore the reliability of Random Forest for resource estimation. This research demonstrates the potential of predictive analytics in advancing cloud resource management, contributing to more efficient and scalable cloud computing practices.**

Keywords—Cloud computing; virtual machine optimization; resource allocation; machine learning

I. Introduction

Cloud computing and virtualization technologies have revolutionized modern computing, offering organizations significant advantages in terms of flexibility, scalability, and operational efficiency [1]. By enabling seamless access to applications and data through online platforms, these technologies ensure constant and universal availability [2]. This has facilitated remote work, improved collaboration among geographically dispersed teams, and accelerated responses to dynamic customer needs, making them indispensable for modern enterprises [3].

The proliferation of cloud technologies has led to significant transformations in IT infrastructure [4]. Emerging paradigms such as hybrid clouds, edge computing, and serverless architectures are redefining how resources are provisioned and utilized. These innovations promise greater adaptability to workload demands but simultaneously introduce complexities in managing and predicting resource needs effectively.

Resource utilization remains one of the most pressing challenges in cloud environments. Infrastructure is often over-provisioned to accommodate peak demands, leading to inefficiencies and inflated costs. Conversely, underutilized resources represent wasted computational potential, underscoring the need for dynamic and predictive strategies to balance workloads effectively [5]. Addressing this challenge is critical for optimizing costs and meeting performance expectations in competitive industries.

Sustainability has also become a pivotal consideration in cloud computing. Data centers are among the most energy-intensive facilities globally, contributing significantly to carbon emissions [6]. Optimizing resource allocation can reduce energy consumption, enabling organizations to align their operations with environmental sustainability goals. These efforts are increasingly essential as industries strive to meet regulatory standards and societal expectations for greener technologies.

Another key challenge is the high operational cost associated with cloud services. Consumption-based pricing models, coupled with additional charges for storage and bandwidth, can complicate budget management, particularly for organizations with fluctuating workloads [7]. Without effective strategies, these financial burdens can hinder the full adoption and utilization of cloud services.

Suboptimal application performance further exacerbates these challenges [8]. Factors such as resource contention among virtual machines (VMs) [9], network latency, and inefficiencies in resource management negatively impact user experience and productivity. These issues can lead to service interruptions, extended downtimes, and reduced competitiveness. Identifying and addressing performance bottlenecks is essential for maintaining application reliability and responsiveness [10].

Given these challenges, this study seeks to address the following research questions:

- How can real-time monitoring and predictive modeling enhance resource allocation in cloud environments?
- What impact do machine learning-based predictive models have on improving cloud resources utilization efficiency?
- How can dynamic resource allocation strategies reduce costs while maintaining optimal system performance?

Based on these questions, the main objectives of this research are:

- To develop a framework that integrates real-time monitoring with predictive modeling to enhance resource efficiency.

- To evaluate the effectiveness of machine learning-based predictive models in improving CPU utilization and system reliability.
- To design and validate dynamic resource allocation techniques that balance workload demand while minimizing costs.

In our previous study [11], we identified CPU utilization as a critical area for improving operational efficiency. However, the limitations in predictive accuracy highlighted the need for more advanced methodologies. Building on these findings, this study presents a comprehensive framework that addresses CPU utilization while tackling broader challenges related to resource allocation, cost management, and system performance. By integrating real-time monitoring, machine learning-based predictive modeling, and dynamic resource allocation techniques, the proposed framework seeks to optimize resource efficiency, reduce costs, and enhance system reliability.

This research provides data-driven strategies that adapt to workload fluctuations, improving both resource utilization and performance. It emphasizes proactive measures to address inefficiencies and enhance cloud systems, contributing to sustainable and scalable cloud computing practices.

The remainder of this article reviews related literature in Section II, details the methodology for data collection and analysis in Section III, Section IV presents the results, and discussion. Finally, the paper is concluded in Section V.

II. Related Work

VM optimization is crucial for enhancing resource utilization and performance in cloud computing environments. Numerous researchers have developed various optimization algorithms and techniques to address this challenge.

Zheng, Huang, Li, and Wang [12] proposed a Cloud Resource Prediction and Migration Method specifically designed for container-based environments. By leveraging machine learning to predict resource demands, their method implemented a migration strategy to balance workloads across containers, thereby improving system performance. Although their work centers on container systems, it provides valuable insights into predictive modeling that can be extended to VM environments.

Kumawat, Handa, and Kharbanda [13] presented a framework for cloud resource optimization tailored for content processing platforms. Using Decision Tree Regression, their approach dynamically assigned instance types based on predicted resource needs, demonstrating the effectiveness of predictive modeling in resource management. However, their work was limited to specific applications, unlike broader approaches applicable across diverse VM workloads.

Shen and Chen [14] developed a Resource-Efficient Predictive Provisioning System for cloud environments. This system utilized resource demand forecasting to optimize allocation and prevent over-provisioning. Their work provides a general framework for improving resource efficiency, but its emphasis is on provisioning rather than VM-specific optimization.

Abbas et al. [15] proposed an ANN-based bidding strategy for resource allocation in cloud computing, utilizing a double auction framework to optimize pricing for IoT applications. Their findings underscored the accuracy of ANN in predicting resource demands and highlighted its potential in complex cloud markets.

Ariza, Jimeno, Villanueva-Polanco, and Capacho [16] applied deep learning models for provisioning resources in cloud-based e-learning platforms. Their approach predicted CPU and memory usage based on real-world data, illustrating how predictive modeling can efficiently adjust resource allocations in response to dynamic demands.

In a related study, Han, Schooley, Mackenzie, David, and Lloyd [17] investigated resource contention in multi-tenant cloud environments. By employing Random Forest models, they predicted resource contention caused by co-located VMs and proposed strategies to mitigate performance degradation. Their study supports the application of machine learning in optimizing VM resource allocation.

Huang, Costero, Pahlevan, Zapater, and Atienza [18] developed CloudProphet, a machine learning-based tool for predicting performance in public cloud environments. By identifying metrics closely correlated with VM performance, their work emphasized the importance of accurate metric selection for resource management.

Wiesi et al. [19] contributed to cloud optimization by using machine learning models such as GRU, LSTM, and Random Forest to predict workloads in dynamic and seasonal environments. Their findings highlighted the role of precise forecasting in improving resource utilization and sustainability.

Ndayikengurukiye et al. [20] proposed the Multi-Objective Seagull Optimization Algorithm Virtual Machine Placement (MOSOAVMP) to optimize VM placement in cloud data centers. Their approach focused on reducing energy consumption, resource wastage, and SLA violations while improving overall efficiency. Simulation results demonstrated significant performance gains over state-of-the-art algorithms, highlighting the effectiveness of this bio-inspired approach for multi-objective optimization.

Another significant contribution comes from Zhang et al. [21] introduced an Extended Coupled Hidden Markov Model (ECHMM) for predicting resource requirements by analyzing historical monitoring data and resource correlations. Although their work focuses on resource prediction, its application in real-time VM optimization remains an open area for exploration.

Building upon these contributions, our study integrates real-time monitoring tools (Prometheus and Grafana) with predictive modeling techniques such as Random Forest and ANN to address gaps in VM optimization. Unlike prior studies, our approach emphasizes dynamic adaptability to changing workloads while achieving significant accuracy improvements for CPU and memory utilization. This integration provides a scalable and efficient solution for resource management across diverse cloud applications, offering valuable insights for adaptive cloud infrastructure management.

III. Approach and Methodology

Our approach to managing and optimizing VM resources combines real-time performance monitoring, data processing, and predictive modeling. The proposed methodology ensures efficient resource utilization, minimizes performance bottlenecks, and reduces operational costs by integrating advanced machine learning techniques with robust monitoring and storage solutions.

A. System Architecture

The system architecture, illustrated in Fig. 1, consists of three main components: VM Monitoring, Data Export and Storage, and Predictive Modeling. These components work in unison to provide real-time insights, store historical data, and enable accurate forecasting for proactive resource allocation.

The first component, VM Monitoring, involves collecting real-time performance metrics such as CPU usage, memory utilization, disk I/O, and network traffic using Prometheus [22], an open-source system designed for time-series data collection. Prometheus integrates seamlessly with the PROXMOX VE virtualization platform, enabling the continuous collection of VM metrics [23]. To visualize this data, Grafana is employed, offering customizable dashboards that provide actionable insights into usage patterns, bottlenecks, and anomalies [24]. This setup ensures proactive resource management by enabling administrators to monitor performance trends in real-time.

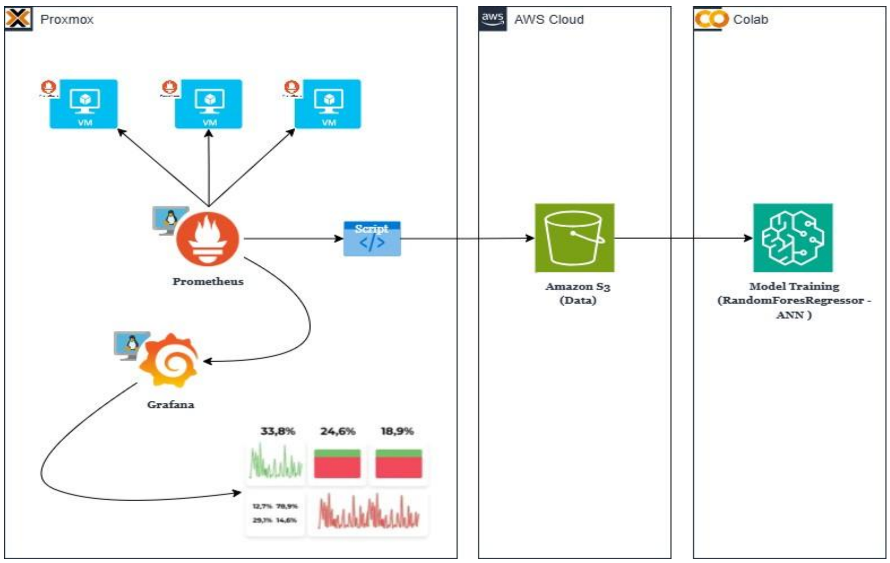

Fig. 1. Resource management and optimization architecture.

To ensure data persistence and facilitate further analysis, performance metrics collected by Prometheus were periodically exported to Amazon S3, a reliable and scalable cloud-based storage solution [25]. Automated scripts managed this process, ensuring reliable backups and accessibility for predictive modeling tasks. IAM policies were applied to secure access to the stored data, which forms the foundation for forecasting future resource demands and optimizing resource allocation strategies.

The final component, Predictive Modeling, leverages machine learning models to analyze historical performance data and anticipate future resource utilization. This enables informed decisions for VM configuration adjustments, ensuring efficient resource usage and avoiding performance bottlenecks.

B. Predictive Modeling

Predictive modeling lies at the core of this methodology, enabling accurate resource demand forecasting to optimize VM allocation. Two techniques were utilized: Random Forest Regression and LSTM networks, each chosen for their robustness and ability to handle complex, nonlinear relationships in resource usage patterns.

Random Forest Regression is an ensemble learning method that combines multiple decision trees [26]. The algorithm creates several decision trees, each trained on a random subset of the data, and aggregates their predictions to produce the final output [27].

This approach reduces overfitting and captures complex interdependencies among variables. Hyperparameters such as the number of trees and maximum depth were fine-tuned to balance prediction accuracy and computational efficiency.

LSTM Networks, a type of recurrent neural network, are designed to capture temporal dependencies in sequential data [28] [29]. They process time-series data by utilizing memory cells with three gates: the Forget Gate, which determines which information to discard; the Input Gate, which decides what new information to incorporate; and the Output Gate, which regulates the information passed to the next layer.

LSTM networks excel at modeling long-term dependencies, making them particularly suited for time-series data such as VM performance metrics. Hyperparameters such as the number of hidden units and learning rate were optimized to ensure accuracy and computational efficiency.

To evaluate the predictive models, two standard metrics were employed: Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). MSE measures the average squared difference between actual and predicted values, as given by:

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

where y_i and $\hat{y}_i$ are the actual and predicted values, respectively, and n is the number of observations. MAPE normalizes the error as a percentage, allowing for a comparative assessment across different scenarios:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

These metrics provided robust insights into the precision and reliability of the models, ensuring their effectiveness in VM resource optimization.

IV. Results and Discussion

The performance of the Random Forest Regression and LSTM models was evaluated for predicting CPU usage in VM. Key performance metrics, including MSE and MAPE, were assessed alongside visual comparisons using forecasting, distribution, scatter, and residual plots.

TABLE I. MODEL PERFORMANCE

Model	MSE	MAPE
Random Forest Regression	0.0011	2.65%
LSTM	0.0137	17.43%

Table I summarizes the comparative performance of the Random Forest and LSTM models. The metrics highlight the superior accuracy of the Random Forest model in predicting CPU usage, as reflected in its lower MSE and significantly better MAPE, demonstrating its robustness for resource optimization tasks.

The Random Forest model demonstrated exceptional alignment between actual and predicted CPU usage, as shown in the forecasting plot (Fig. 2). The predicted values closely tracked the observed data, even during abrupt transitions, showcasing the model's ability to adapt to workload fluctuations. The distribution plot (Fig. 4) revealed that the predicted values nearly overlapped with the actual distribution, confirming the model's precision in capturing data variability. The scatter plot (Fig. 6) further substantiated these findings, with data points tightly clustered along the diagonal, indicating minimal prediction errors and a high correlation between actual and predicted values. Moreover, the residual plot (Fig. 8) presented a near-uniform distribution centered around zero, reflecting the model's unbiased performance and robust generalization across diverse workloads.

In contrast, the LSTM model exhibited noticeable discrepancies. The forecasting plot (Fig. 3) showed that while the model successfully captured general trends in CPU usage, its performance during abrupt changes was suboptimal, with evident prediction lags. The distribution plot (Fig. 5) illustrated significant deviations between actual and predicted values, with broader peaks and a misaligned density curve, suggesting difficulties in modeling the variability and complexity of resource utilization patterns. The scatter plot (Fig. 7) highlighted this challenge further, with pronounced dispersion away from the diagonal, signifying higher prediction errors, particularly under extreme workload conditions. The residual plot (Fig. 9) revealed non-random patterns, with clusters of over- and underprediction, pointing to biases in the model's predictions and underscoring the need for further tuning and optimization.

Fig. 2. Forecasting time series plot (Random forest).

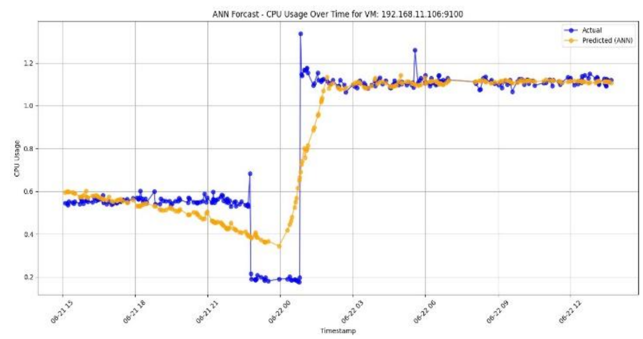

Fig. 3. Forecasting time series plot (LSTM).

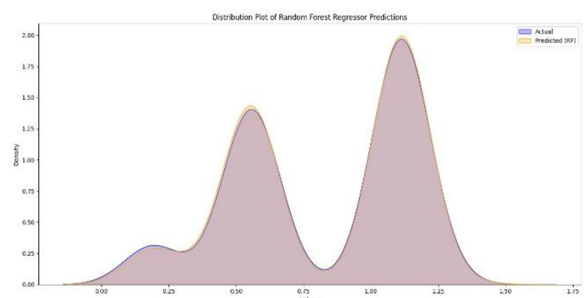

Fig. 4. Distribution plot (Random forest).

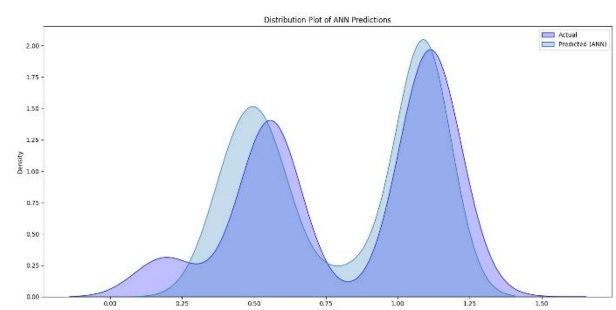

Fig. 5. Distribution plot (LSTM).

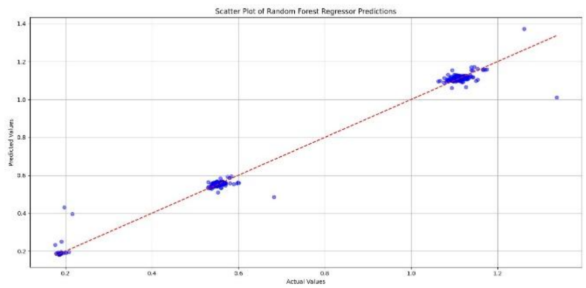

Fig. 6. Scatter plot (Random forest).

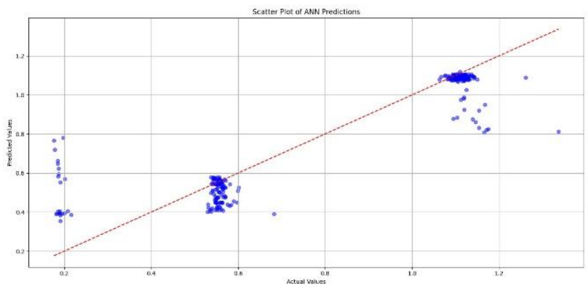

Fig. 7. Scatter plot (LSTM).

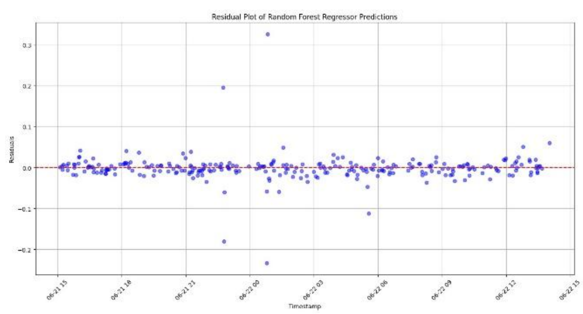

Fig. 8. Residual plot (Random forest).

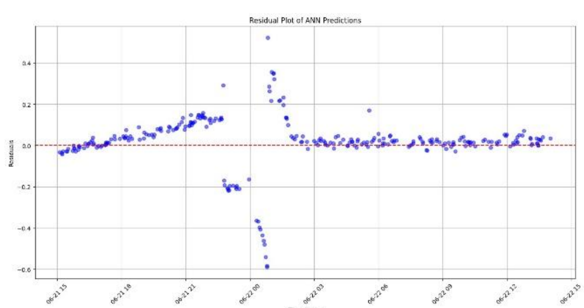

Fig. 9. Residual plot (LSTM).

Additionally, the residual analysis (Fig. 5) of the Random Forest model indicated greater robustness and reliability in handling varying workloads. In contrast, the LSTM model's underperformance highlighted the challenges of applying deep learning to dynamic workloads without substantial hyperparameter tuning and larger datasets for training.

This study's integration of predictive models for VM optimization builds upon and extends existing research. In comparison to Shen and Chen [14], who developed a provisioning system for general resource allocation, this work incorporates VM-specific adaptability through real-time monitoring. Shen and Chen's approach lacks the dynamic allocation capabilities achieved here, as reflected in the superior MAPE of 2.65% obtained by the Random Forest model.

Similarly, Abbas et al. [15] employed an ANN-based bidding strategy for IoT resource pricing, achieving high accuracy for specific applications. However, the computational complexity of ANN models limits their broader applicability. In contrast, the Random Forest model balances accuracy and efficiency, making it more practical for general VM optimization.

The container-based optimization approaches of Zheng et al. [12] and content-specific frameworks of Kumawat et al. [13] demonstrate effective solutions for narrow contexts but lack the versatility of this study's framework, which dynamically adapts to diverse workloads using real-time monitoring tools like Prometheus and Grafana. Moreover, while Han et al. [17] addressed resource contention using Random Forest models, their focus on co-residency prediction differs from this study's broader goal of optimizing resource utilization and preventing performance degradation.

Additionally, Huang et al. [18] developed CloudProphet, a performance prediction tool for public clouds, which focuses on general workload trends but lacks real-time data integration. This study's use of Prometheus for live data collection ensures greater adaptability and real-world applicability, particularly under dynamic conditions. The seasonal workload prediction by Wiesi et al. [19], which used GRU and LSTM models, also does not address abrupt workload changes, further highlighting the Random Forest model's robustness in handling dynamic scenarios.

This study also represents a significant improvement over our previous research. In the earlier work, simpler modeling techniques and less dynamic monitoring systems were used, leading to a MAPE of 11% for CPU utilization predictions. By incorporating real-time monitoring tools, such as Prometheus and Grafana, alongside advanced predictive modeling with Random Forest, this study reduced the MAPE to 2.65%. This fourfold improvement in predictive accuracy reflects the effectiveness of the enhanced framework in addressing the limitations identified in the earlier study. The integration of real-time monitoring and advanced ensemble learning has enabled the system to capture more complex resource utilization patterns, offering a more reliable and adaptive solution for VM optimization.

The observed differences between the two models provide critical insights. The Random Forest model's ensemble approach, which aggregates predictions from multiple decision trees, allows it to effectively balance accuracy and generalization. This characteristic is particularly advantageous in resource optimization scenarios where precision is vital. On the other hand, the LSTM model, while less accurate, has potential for scalability and adaptability in handling larger datasets or real-time applications if further refined. Its performance limitations in this study emphasize the need for hybrid modeling approaches that combine statistical methods and neural networks to enhance predictive accuracy.

The findings have significant implications for real-world applications. The superior performance of the Random Forest model makes it a reliable choice for real-time resource optimization in cloud computing environments, where accurate predictions are essential for cost efficiency and service quality. However, the LSTM model's potential scalability and adaptability warrant further exploration, particularly in scenarios with high variability and evolving workloads.

Despite its advantages, this study has certain limitations. The accuracy of the predictive model relies on the quality and consistency of real-time monitoring data. Variations in cloud workload patterns may also introduce unexpected challenges, potentially affecting resource allocation precision.

This study highlights the importance of selecting appropriate modeling techniques for resource prediction in cloud environments. By demonstrating the strengths of ensemble learning and identifying the limitations of deep learning in this context, the research provides a foundation for future work. Subsequent studies could explore the integration of hybrid models or the application of advanced deep learning architectures to improve predictive performance. Furthermore, real-world deployment of these models in diverse cloud infrastructures will validate their practical utility and scalability, contributing to more efficient resource management and optimization.

V. Conclusion

This research presents a comprehensive framework for optimizing VM performance within cloud computing environments by integrating advanced machine learning methodologies and real-time monitoring tools. The study highlights the exceptional efficacy of the Random Forest Regression model, which achieved a MAPE of 2.65%, significantly outperforming the LSTM model. This reduction in prediction error underscores the model's ability to enable precise resource allocation, leading to substantial improvements in system performance, operational efficiency, and cost-effectiveness.

Compared to traditional approaches and prior studies, this framework represents a critical advancement in cloud resource management. The integration of real-time monitoring tools, such as Prometheus and Grafana, combined with advanced predictive analytics, enables dynamic adaptability to workload changes and more efficient resource utilization. By addressing key limitations of earlier research, such as reliance on less adaptive systems or single-method approaches, this study establishes a new benchmark for VM optimization, particularly by demonstrating the robustness of ensemble learning techniques like Random Forest in handling complex and dynamic resource utilization patterns.

While this research demonstrates significant progress, future work could explore the integration of the Random Forest model into a hybrid framework, building upon the strengths of ensemble learning and deep learning. Such an approach could leverage the Random Forest model's robust accuracy alongside LSTM's ability to handle sequential patterns, creating a scalable and adaptive solution for even more complex cloud environments. Additionally, extending the framework to incorporate metrics like energy consumption and applying it across multi-cloud environments would further enhance its utility and relevance. Real-world deployment and validation in diverse cloud infrastructures will be essential for solidifying its practical impact and scalability. In conclusion, the proposed framework offers a versatile and practical solution for addressing the challenges of modern cloud computing environments, paving the way for more efficient and sustainable cloud operations.

References

[1] O. Obi, S. Dawodu, A. Daraojimba, S. Onwusinkwue, O. Akagha, and I. Ahmad, “REVIEW OF EVOLVING CLOUD COMPUTING PARADIGMS: SECURITY, EFFICIENCY, AND INNOVATIONS,” Computer Science & IT Research Journal, vol. 5, pp. 270–292, Feb. 2024, doi: 10.51594/csitrj.v5i2.757.

[2] Y. Wang, Q. Bao, J. Wang, G. Su, and X. Xu, “Cloud Computing for Large-Scale Resource Computation and Storage in Machine Learning,” JTPES, vol. 4, no. 03, pp. 163–171, Mar. 2024, doi: 10.53469/jtpes.2024.04(03).14.

[3] M. Attaran, S. Attaran, and D. Kirkland, “Technology and Organizational Change: Harnessing the Power of Digital Workplace,” 2019, pp. 383–408. doi: 10.4018/978-1-5225-8933-4.

[4] M. E. E. Alahi et al., “Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends,” Sensors, vol. 23, no. 11, Art. no. 11, Jan. 2023, doi: 10.3390/s23115206.

[5] P. K. G. Pandian, “Effective Resource Management In Virtualized Environments,” vol. 1, no. 7, 2023.

[6] M. Yenugula, S. Sahoo, and S. Goswami, “Cloud computing for sustainable development: An analysis of environmental, economic and social benefits,” Journal of Future Sustainability, vol. 4, no. 1, pp. 59–66, 2024.

[7] R. Islam et al., “The Future of Cloud Computing: Benefits and Challenges,” International Journal of Communications, Network and System Sciences, vol. 16, no. 4, Art. no. 4, Apr. 2023, doi: 10.4236/ijcns.2023.164004.

[8] H. Ahmed, H. J. Syed, A. Sadiq, A. O. Ibrahim, M. Alohaly, and M. Elsadig, “Exploring Performance Degradation in Virtual Machines Sharing a Cloud Server,” Applied Sciences, vol. 13, no. 16, Art. no. 16, Jan. 2023, doi: 10.3390/app13169224.

[9] S. Kraft, G. Casale, D. Krishnamurthy, D. Greer, and P. Kilpatrick, “Performance models of storage contention in cloud environments,” Softw Syst Model, vol. 12, no. 4, pp. 681–704, Oct. 2013, doi: 10.1007/s10270-012-0227-2.

[10] Y. Gong, J. Huang, B. Liu, J. Xu, B. Wu, and Y. Zhang, “Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning,” Mar. 20, 2024, arXiv: arXiv:2403.13619. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/2403.13619

[11] R. Doukha, A. Ez-Zahout, and A. Ndayikengurukiye, “Forecasting virtual machine resource utilization in cloud computing: a hybrid artificial intelligence approach,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 37, no. 3, Art. no. 3, Mar. 2025, doi: 10.11591/ijeecs.v37.i3.pp1887-1898.

[12] S. Zheng, F. Huang, C. Li, and H. Wang, “A Cloud Resource Prediction and Migration Method for Container Scheduling,” in 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dec. 2021, pp. 76–80. doi: 10.1109/TOCS53301.2021.9689034.

[13] N. Kumawat, N. Handa, and A. Kharbanda, “Cloud Computing Resources Utilization and Cost Optimization for Processing Cloud Assets,” in 2020 IEEE International Conference on Smart Cloud (SmartCloud), Washington DC, WA, USA: IEEE, Nov. 2020, pp. 41–48. doi: 10.1109/SmartCloud49737.2020.00017.

[14] H. Shen and L. Chen, “A Resource-Efficient Predictive Resource Provisioning System in Cloud Systems,” IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 12, pp. 3886–3900, Dec. 2022, doi: 10.1109/TPDS.2022.3172493.

[15] M. Adeel Abbas, Z. Iqbal, F. Zeeshan Khan, S. Alsubai, A. Binbusayyis, and A. Alqahtani, “An ANN based bidding strategy for resource allocation in cloud computing using IoT double auction algorithm,” Sustainable Energy Technologies and Assessments, vol. 52, p. 102358, Aug. 2022, doi: 10.1016/j.seta.2022.102358.

[16] J. Ariza, M. Jimeno, R. Villanueva-Polanco, and J. Capacho, “Provisioning Computational Resources for Cloud-Based e-Learning Platforms Using Deep Learning Techniques,” IEEE Access, vol. PP, pp. 1–1, Jun. 2021, doi: 10.1109/ACCESS.2021.3090366.

[17] X. Han, R. Schooley, D. Mackenzie, O. David, and W. J. Lloyd, “Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction,” in 2020 IEEE International Conference on Cloud Engineering (IC2E), Apr. 2020, pp. 162–172. doi: 10.1109/IC2E48712.2020.00024.

[18] D. Huang, L. Costero, A. Pahlevan, M. Zapater, and D. Atienza, “CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds,” Sep. 28, 2023, arXiv: arXiv:2309.16333. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/2309.16333

[19] A. K. Wiesi et al., “Optimizing Cloud Resource Utilization Through Machine Learning Forecasting,” . Vol., no. 17.

[20] A. Ndayikengurukiye, R. Doukha, E. Niyukuri, E. Muheto, A. Ez-zahout, and F. Omary, “SOAVMP: Multi-Objective Virtual Machine Placement in Cloud Computing Based on the Seagull Optimization Algorithm,” IJCNA, vol. 11, no. 3, p. 375, Jun. 2024, doi: 10.22247/ijcna/2024/24.

[21] J. He, S. Hong, C. Zhang, Y. Liu, F. Deng, and J. Yu, “A Method to Cloud Computing Resources Requirement Prediction on SaaS Application,” in 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Jul. 2021, pp. 107–116. doi: 10.1109/MLISE54096.2021.00027.

[22] J. Turnbull, Monitoring with Prometheus. Turnbull Press, 2018.

[23] S. A. Algarni, M. R. Ikbal, R. Alroobaea, A. S. Ghiduk, and F. Nadeem, “Performance Evaluation of Xen, KVM, and Proxmox Hypervisors,” IJOSSP, vol. 9, no. 2, pp. 39–54, Apr. 2018, doi: 10.4018/IJOSSP.2018040103.

[24] M. Chakraborty and A. P. Kundan, “Grafana,” in Monitoring Cloud-Native Applications: Lead Agile Operations Confidently Using Open Source Software, M. Chakraborty and A. P. Kundan, Eds., Berkeley, CA: Apress, 2021, pp. 187–240. doi: 10.1007/978-1-4842-6888-9_6.

[25] S. Gulabani, Amazon S3 Essentials. Packt Publishing Ltd, 2015.

[26] R. M. Schulte, M. D. Lebsock, J. M. Haynes, and Y. Hu, “A random forest algorithm for the prediction of cloud liquid water content from combined CloudSat–CALIPSO observations,” Atmospheric Measurement Techniques, vol. 17, no. 11, pp. 3583–3596, Jun. 2024, doi: 10.5194/amt-17-3583-2024.

[27] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, “A comparison of random forest variable selection methods for classification prediction modeling,” Expert Systems with Applications, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[28] A. Gozuoglu, O. Ozgonenel, and C. Gezegin, “CNN-LSTM based deep learning application on Jetson Nano: Estimating electrical energy consumption for future smart homes,” Internet of Things, vol. 26, p. 101148, Jul. 2024, doi: 10.1016/j.iot.2024.101148.

[29] J. Wang, S. Hong, Y. Dong, Z. Li, and J. Hu, “Predicting Stock Market Trends Using LSTM Networks: Overcoming RNN Limitations for Improved Financial Forecasting,” Journal of Computer Science and Software Applications, vol. 4, no. 3, pp. 1–7, Jul. 2024, doi: 10.5281/zenodo.12200708.