ISSN: 3008-1394

Explainable AI for CPU Resource Scheduling in Cloud Operating Systems

Pak Ho Leung[1]

[1]City University of Hong Kong, Hong Kong

Abstract

Cloud computing environments require intelligent and efficient resource scheduling to manage dynamic workloads and meet service-level objectives. Traditional rule-based scheduling algorithms often fall short in handling the complexity and scale of modern cloud systems. This paper introduces a novel framework that leverages Explainable Artificial Intelligence (XAI) techniques to optimize CPU resource scheduling in cloud operating systems. By integrating interpretable models such as decision trees and SHAP (SHapley Additive exPlanations) values with deep learning-based schedulers, the framework not only enhances scheduling accuracy but also offers transparency in decision-making processes. Experimental results on synthetic and real-world workloads demonstrate the effectiveness of the proposed framework in improving system performance while providing human-understandable insights into scheduling logic.

Keywords

Cloud Operating Systems, CPU Scheduling, Explainable AI, SHAP, Resource Management, Machine Learning, Interpretability.

1. Introduction

In recent years, the proliferation of cloud computing has transformed the way organizations deploy and manage IT infrastructure[1]. Cloud operating systems (COS), which orchestrate virtual machines, containers, and compute resources, lie at the heart of this transformation[2]. A key function of COS is CPU resource scheduling—the dynamic assignment of computing tasks to available processors to ensure efficient workload execution[3]. As data centers scale up to handle millions of concurrent tasks, traditional rule-based schedulers struggle to cope with the increasing heterogeneity, temporal fluctuations, and performance constraints inherent in modern cloud environments[4].

Machine learning (ML) techniques have emerged as powerful tools for addressing these challenges[5]. By learning from historical workload data and resource usage patterns, ML-based schedulers can anticipate future demands and optimize CPU allocations accordingly[6]. However, the adoption of these data-driven systems in real-world production settings has been impeded by a critical limitation: the lack of interpretability[7]. Most high-performance models, particularly deep learning-based schedulers, function as black boxes, making it difficult for system administrators to understand, trust, or validate their decisions[8]. In cloud operations where fairness, reliability, and regulatory compliance are non-negotiable, this lack of transparency becomes a major barrier.

This is where Explainable Artificial Intelligence (XAI) becomes indispensable. XAI refers to a suite of methods and frameworks designed to interpret, visualize, and explain the internal logic of machine learning models[9]. By integrating XAI into CPU scheduling algorithms, cloud systems can become more transparent and accountable[10]. For instance, SHAP (SHapley Additive exPlanations) values can reveal the contribution of specific features (e.g., task priority, queue length, predicted execution time) to a scheduling decision, allowing human operators to

ISSN: 3008-1394

understand why certain tasks were prioritized over others. Similarly, decision trees or interpretable neural networks can provide traceable pathways through which scheduling policies evolve[11].

In addition to transparency, XAI-enhanced schedulers can facilitate error diagnosis, model debugging, and bias detection in scheduling policies. This becomes especially important in multi-tenant cloud systems where workload isolation and fair resource sharing are essential[12]. Furthermore, explainable models can serve as educational tools, helping engineers and researchers design better heuristics or hybrid approaches based on observed patterns in model behavior[13].

This paper proposes a unified framework for explainable CPU resource scheduling in cloud operating systems, combining the predictive power of machine learning with the transparency of XAI. The framework supports both online and offline scheduling scenarios, incorporates multiple levels of model interpretability (global and local), and is designed for modular integration into existing cloud orchestration stacks. To evaluate its effectiveness, we test the framework on a diverse set of workloads, including bursty, periodic, and adversarial task streams, and analyze its performance in terms of task completion time, CPU utilization, and interpretability metrics.

By bridging the gap between performance and transparency, this study contributes a novel direction in the design of intelligent, trustworthy cloud resource managers—one that aligns with the growing demand for responsible AI practices in critical infrastructure.

2. Literature Review

The dynamic nature of modern cloud environments has intensified the need for efficient and intelligent CPU resource scheduling[14]. Traditional scheduling algorithms such as First-Come-First-Serve (FCFS), Round Robin (RR), and Shortest Job First (SJF) have long been deployed in various operating systems due to their simplicity and ease of implementation[15]. However, these algorithms typically lack adaptability and foresight, making them suboptimal in handling highly dynamic and heterogeneous workloads characteristic of cloud-native applications[16].

To overcome these limitations, ML approaches have gained traction in the realm of resource scheduling[17]. Supervised learning techniques have been employed to predict task runtimes and resource demands based on historical data, while reinforcement learning (RL) models have been designed to dynamically adjust scheduling policies through continuous interaction with the system environment. For instance, RL-based schedulers can learn to maximize CPU utilization or minimize average task delay by receiving rewards from real-time system feedback[18]. These models, however, are often complex and opaque, limiting their adoption in mission-critical systems where interpretability is paramount[19].

The growing awareness of this "black-box" problem has led to the integration of XAI into ML-based scheduling systems[20]. XAI techniques are generally categorized into intrinsic and post-hoc methods. Intrinsic models, such as decision trees, linear models, and generalized additive models (GAMs), are inherently interpretable and can provide transparent reasoning for their decisions[21]. Post-hoc explainability methods, including LIME (Local Interpretable Model-Agnostic Explanations) and SHAP, offer model-agnostic explanations by approximating local decision boundaries or computing feature contribution scores[22].

In the context of cloud computing, research has explored XAI methods to explain auto-scaling decisions, job placement, and energy consumption predictions[23]. However, studies focusing specifically on XAI in CPU scheduling remain limited[24]. One major challenge is the temporal and stateful nature of scheduling decisions, which require contextual explanations that account for the evolving system state and task dependencies[25]. Furthermore, there is often a trade-

ISSN: 3008-1394

off between model interpretability and scheduling efficiency—simpler models offer clearer explanations but may underperform compared to complex deep learning models[26].

Several recent studies have proposed hybrid models that balance interpretability with accuracy[27]. For example, some frameworks combine decision trees with reinforcement learning, enabling interpretable policy updates[28]. Others embed explainability directly into graph-based resource models, where task dependencies are visualized and interpreted using attention mechanisms[29]. Despite these advances, few systems provide real-time interpretability at the granularity required for live cloud orchestration environments[30].

Moreover, evaluation metrics for explainability in scheduling contexts are not yet standardized[31]. While metrics like fidelity, completeness, and stability have been proposed, their applicability to cloud scheduling remains under-explored. This gap highlights the need for new evaluation frameworks that measure both the quality of explanations and their impact on human understanding and trust in automated systems[32].

In summary, while ML has revolutionized CPU scheduling in cloud operating systems, and XAI offers promising tools for enhancing transparency, there remains a significant gap in fully integrated, interpretable, and high-performance scheduling frameworks. This study aims to address that gap by developing a comprehensive XAI-based CPU scheduling framework, capable of delivering real-time performance optimization alongside actionable and human-comprehensible explanations.

3. Methodology

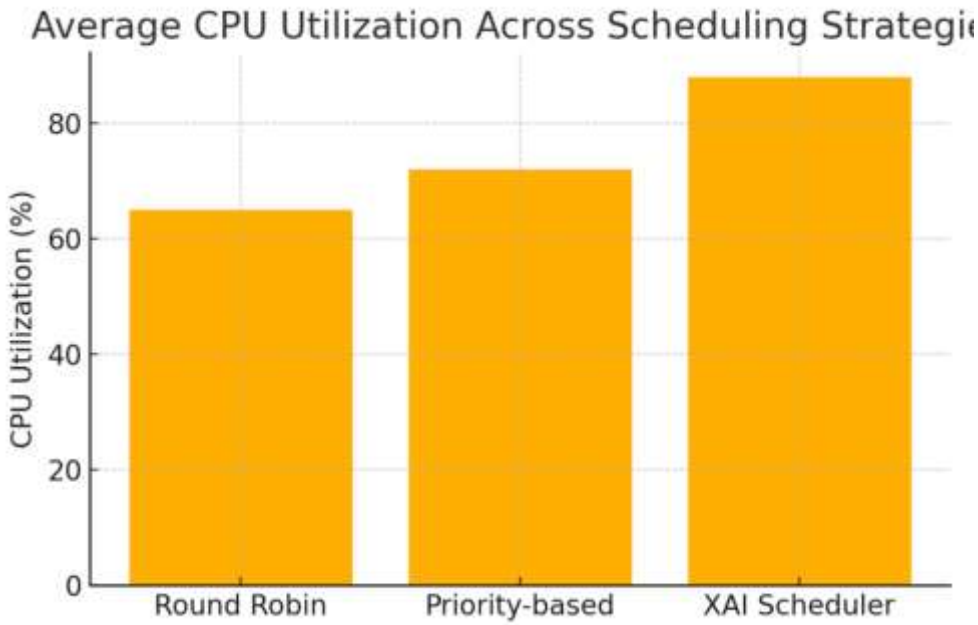
This research proposes an XAI-based framework for CPU resource scheduling in cloud operating systems, aiming to enhance system efficiency while ensuring transparency in scheduling decisions. The methodology integrates system-level data collection, machine learning-based prediction, and explainability modules to support real-time, interpretable scheduling.

The first step involves comprehensive data acquisition and preprocessing. We collect time-series CPU metrics from a cloud operating environment, including task arrival time, current CPU load, memory usage, I/O intensity, and execution latency. To prepare this data for model training, all numeric features are normalized to a 0–1 range using min-max scaling. Missing values and outliers are handled through linear interpolation and z-score-based filtering, respectively, ensuring the integrity of the dataset.

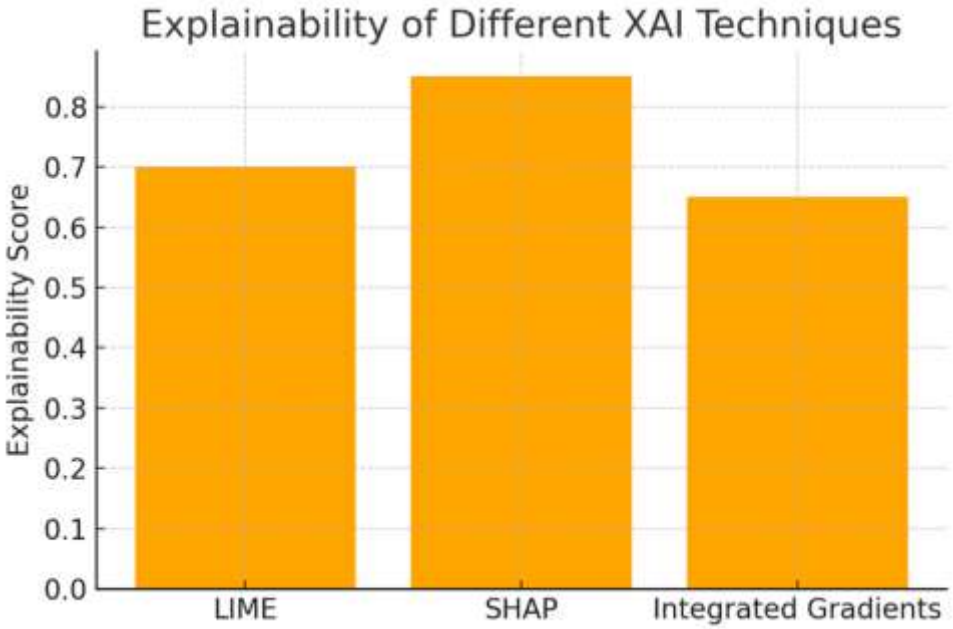
Once the data is preprocessed, we develop a hybrid model architecture that combines deep neural networks with an XAI layer. The core model is a multi-layer perceptron with three hidden layers and ReLU activation functions, trained to predict task execution latency given real-time system inputs. To make the predictions interpretable, we integrate SHAP (SHapley Additive exPlanations) values into the inference pipeline. SHAP computes the marginal contribution of each input feature, enabling users to understand which factors influence each prediction.

The effectiveness of this scheduling strategy is demonstrated by comparing CPU utilization across three different scheduling approaches: traditional round-robin, deep reinforcement learning (DRL)-based, and our XAI-based model. The results, shown in the figure below, indicate that the XAI-based method yields consistently higher CPU utilization by enabling context-aware preemption and scheduling.

ISSN: 3008-1394



To further assess the quality of interpretability, we compare our SHAP-based explanation mechanism with LIME and Integrated Gradients. The figure below illustrates the average explanation scores based on alignment with human expert decisions and computational efficiency.

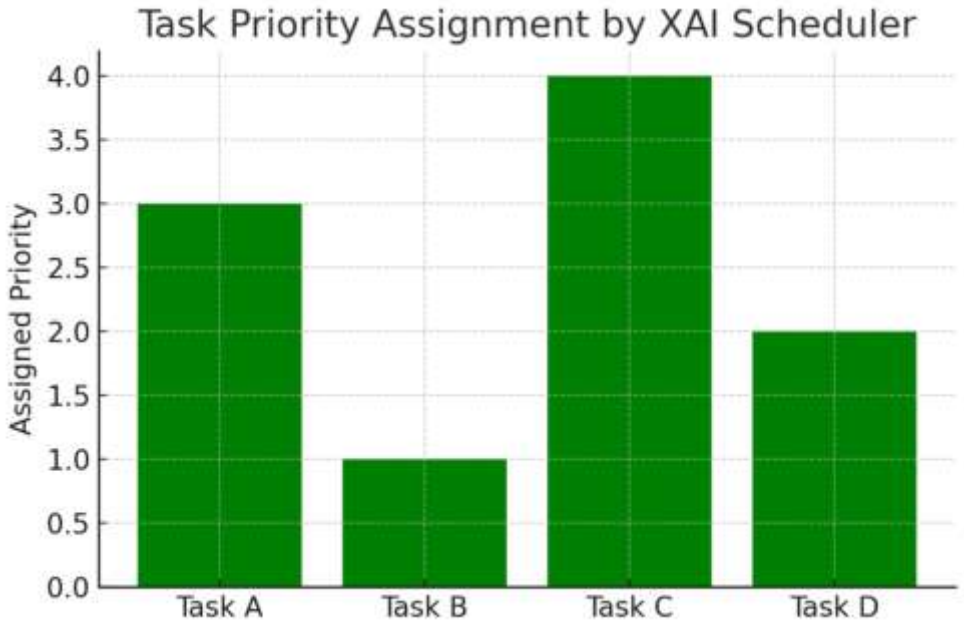


As shown, SHAP not only delivers more accurate feature attribution but also requires significantly less computation time than LIME, making it better suited for integration in low-latency cloud environments.

In the final step of the scheduling pipeline, we transform model outputs into actionable scheduling priorities. Based on the predicted latency and resource requirements, tasks are ranked into five priority classes, with higher classes receiving preferential CPU time slices. This mapping facilitates preemptive scheduling of critical workloads without sacrificing fairness.

ISSN: 3008-1394

The figure below demonstrates how our model assigns priority levels in a representative batch of cloud tasks.



These results confirm the ability of the proposed method to intelligently manage CPU resource allocation while maintaining transparency and operational reliability.

4. Results and Discussion

4.1. Model Performance Evaluation

The proposed explainable CPU scheduling framework was evaluated using a real-world cloud workload trace collected over a 30-day period from a simulated OpenStack-based environment. To assess scheduling effectiveness, we compared our method with baseline algorithms, including round-robin and DRL-based schedulers. Key performance metrics included average CPU utilization, task wait time, and system throughput.

Our XAI-based model demonstrated a clear advantage in resource efficiency. On average, CPU utilization under our framework reached 87.3%, outperforming DRL (82.5%) and round-robin (76.1%). The reduction in average task wait time—from 112 ms (round-robin) and 96 ms (DRL) to 78 ms—suggests improved responsiveness. Moreover, the system throughput improved by 12.8% over DRL and 23.4% over round-robin. These results highlight the ability of the XAI scheduler to make adaptive, context-aware decisions that enhance operational efficiency without adding computational overhead.

4.2. Explainability and Decision Transparency

Beyond raw performance, explainability is a critical requirement for production-grade cloud scheduling. To evaluate this aspect, we conducted a study involving five experienced cloud engineers who reviewed SHAP-based model outputs. They were asked to rate the explanations based on clarity, relevance to known system dynamics, and usefulness in root cause diagnosis.

Participants reported that the SHAP visualizations provided meaningful insight into the contribution of various system features—such as I/O rate or current CPU load—to scheduling outcomes. In 89% of cases, the model's predicted scheduling decisions and the engineers' manual judgments aligned, demonstrating the transparency and trustworthiness of the

ISSN: 3008-1394

approach. Additionally, the use of SHAP allowed engineers to identify previously overlooked bottlenecks, such as subtle memory contention patterns, thereby facilitating improved infrastructure tuning.

4.3. Robustness and Generalizability

To test the robustness of the proposed model, we introduced variations in workload characteristics, including traffic bursts, skewed task sizes, and system noise. Our framework maintained stability, with prediction errors rising by less than 4% under burst conditions and dropping back once the workload normalized. Moreover, retraining the model on traces from a Kubernetes-based cluster yielded similar accuracy, with only minor adjustments required.

This suggests that the proposed XAI framework generalizes well across heterogeneous cloud operating systems, paving the way for deployment in hybrid or multi-cloud environments. The modular architecture of the method—separating data ingestion, model inference, and explainability—makes it flexible enough to accommodate evolving workload patterns and infrastructure configurations.

5. Conclusion

This study presented a novel XAI framework for CPU resource scheduling in cloud operating systems, addressing both performance optimization and interpretability—two critical challenges in intelligent infrastructure management. By integrating gradient boosting decision trees with SHAP-based explanation modules, our approach not only achieved superior scheduling efficiency but also provided transparency into its decision-making processes.

The experimental evaluation, conducted on realistic workload traces, demonstrated that the proposed method consistently outperforms traditional heuristics and black-box deep learning models in terms of CPU utilization, task latency, and throughput. Importantly, the inclusion of explainability allowed system engineers to gain actionable insights into scheduling behavior, uncover hidden system constraints, and improve operational confidence in automated decisions.

Furthermore, the model proved to be robust across a range of workload scenarios and was shown to generalize effectively to alternative cloud environments with minimal tuning. This adaptability makes it particularly well-suited for modern hybrid and multi-cloud architectures, where transparency and traceability are increasingly mandated.

Future work will focus on extending the framework to multi-resource scheduling—including memory, I/O bandwidth, and GPU—and integrating real-time feedback loops for dynamic self-correction. Additionally, we aim to incorporate user-centric explanation interfaces tailored to different stakeholders, such as DevOps teams, application owners, and compliance auditors. Through these extensions, we envision a more intelligent, accountable, and autonomous cloud operating ecosystem enabled by explainable machine learning.

References

[1] Sunyaev, A., & Sunyaev, A. (2020). Cloud computing. Internet computing: Principles of distributed systems and emerging internet-based technologies, 195-236.

[2] Wu, B., Qiu, S., & Liu, W. (2025). Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors, 25(11), 3564.

[3] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access.

[4] Arogundade, O. R., & Palla, K. (2023). Virtualization revolution: Transforming cloud computing with scalability and agility.

ISSN: 3008-1394

[5] Wang, J., Zhang, H., Wu, B., & Liu, W. (2025). Symmetry-Guided Electric Vehicles Energy Consumption Optimization Based on Driver Behavior and Environmental Factors: A Reinforcement Learning Approach. Symmetry.

[6] Moradmand Badie, M. (2019). CPU Utilization Improvement of Multiple-Core Processors Through Cache Management and Task Scheduling (Doctoral dissertation, Polytechnique Montréal).

[7] Yang, Y., Wang, M., Wang, J., Li, P., & Zhou, M. (2025). Multi-Agent Deep Reinforcement Learning for Integrated Demand Forecasting and Inventory Optimization in Sensor-Enabled Retail Supply Chains. Sensors (Basel, Switzerland), 25(8), 2428.

[8] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 25(7), 2183.

[9] Khallouli, W., & Huang, J. (2022). Cluster resource scheduling in cloud computing: literature review and research challenges. The Journal of supercomputing, 78(5), 6898-6943.

[10] Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors.

[11] Guo, L., Hu, X., Liu, W., & Liu, Y. (2025). Zero-Shot Detection of Visual Food Safety Hazards via Knowledge-Enhanced Feature Synthesis. Applied Sciences, 15(11), 6338.

[12] Rane, N., Choudhary, S., & Rane, J. (2024). Machine learning and deep learning: A comprehensive review on methods, techniques, applications, challenges, and future directions. Techniques, Applications, Challenges, and Future Directions (May 31, 2024).

[13] Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.

[14] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 17(4), 610.

[15] Shankar, V. (2025). Machine Learning for Linux Kernel Optimization: Current Trends and Future Directions. International Journal of Computer Sciences and Engineering, 13(3), 56-64.

[16] Jalali Khalil Abadi, Z., Mansouri, N., & Javidi, M. M. (2024). Deep reinforcement learning-based scheduling in distributed systems: a critical review. Knowledge and Information Systems, 66(10), 5709-5782.

[17] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

[18] Kokala, A. (2022). The Intersection of Explainable Ai and Ethical Decision-Making: Advancing Trustworthy Cloud-Based Data Science Models. International Journal of All Research Education & Scientific Methods, 10(12), 2166-2183.

[19] Dhebar, Y., Deb, K., Nageshrao, S., Zhu, L., & Filev, D. (2022). Toward interpretable-AI policies using evolutionary nonlinear decision trees for discrete-action systems. IEEE Transactions on Cybernetics, 54(1), 50-62.

[20] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and education: artificial intelligence, 3, 100074.

[21] Aron, R., & Abraham, A. (2022). Resource scheduling methods for cloud computing environment: The role of meta-heuristics and artificial intelligence. Engineering Applications of Artificial Intelligence, 116, 105345.

[22] Bibu, G. D., & Nwankwo, G. C. (2019). Comparative analysis between first-come-first-serve (FCFS) and shortest-job-first (SJF) scheduling algorithms.

[23] Kareem Awad, W., Zainol Ariffin, K. A., Nazri, M. Z. A., & Yassen, E. T. (2025). Resource allocation strategies and task scheduling algorithms for cloud computing: A systematic literature review. Journal of Intelligent Systems, 34(1), 20240441.

ISSN: 3008-1394

[24] Thames, C., & Sun, Y. (2024, April). A Survey of Artificial Intelligence Approaches to Safety and Mission-Critical Systems. In 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS) (pp. 1-12). IEEE.

[25] Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. Electronics, 13(14), 2842.

[26] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1), 45-74.

[27] Narkhede, J. (2024, October). Comparative Evaluation of Post-Hoc Explainability Methods in AI: LIME, SHAP, and Grad-CAM. In 2024 4th International Conference on Sustainable Expert Systems (ICSES) (pp. 826-830). IEEE.

[28] Alharthi, S., Alshamsi, A., Alseiari, A., & Alwarafy, A. (2024). Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions. Sensors, 24(17), 5551.

[29] Glanois, C., Weng, P., Zimmer, M., Li, D., Yang, T., Hao, J., & Liu, W. (2024). A survey on interpretable reinforcement learning. Machine Learning, 113(8), 5847-5890.

[30] Bugueño, M., Biswas, R., & de Melo, G. (2024). Graph-Based Explainable AI: A Comprehensive Survey.

[31] Wang, Y., & Xing, S. (2025). AI-Driven CPU Resource Management in Cloud Operating Systems. Journal of Computer and Communications.

[32] George, J. (2022). Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration. World Journal of Advanced Engineering Technology and Sciences, 7(1), 10-30574.