IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, January 2026
Impact Factor: 7.67

Review on Quantum ML Workload Prediction and Categorization in Cloud Computing

Mahajan Sanchit Shrikant[1], Dr. Parvathraj K M M[2], Dr. Sunny B. Mohite[3]
PhD Scholar, CSE, Srinivas University, Mangluru, Mukka, Karnataka, India[1]
Professor, AI&ML, Srinivas Institute of Technology, Mangluru, Mukka, Karnataka, India[2]
Professor, CSE, D Y Patil College of Engineering and Technology, Kolhapur, Maharashtra, India[3]

Abstract: *The rapid advancement of cloud computing has enabled large-scale deployment of machine learning applications; however, the emergence of Quantum Machine Learning (QML) introduces new challenges for cloud workload management. QML workloads exhibit unique execution patterns, hybrid quantum–classical behavior, and distinct resource requirements that are not effectively handled by existing cloud workload prediction and scheduling mechanisms. This paper presents a comprehensive review of recent research on cloud workload optimization, scheduling, and characterization, highlighting their limitations in supporting QML workloads. Based on the identified research gaps, a system model and methodology for QML workload prediction and categorization in cloud computing environments are discussed. The proposed approach emphasizes early workload analysis, feature-based prediction, and intelligent categorization to enable proactive scheduling and efficient resource utilization. Furthermore, key challenges related to quantum hardware limitations, hybrid integration complexity, and prediction reliability are analyzed. This study provides foundational insights for designing scalable and intelligent cloud platforms capable of supporting next-generation Quantum Machine Learning applications.*

Keywords: Quantum Machine Learning, Cloud Computing, Workload Prediction, Workload Categorization, Hybrid Quantum–Classical Systems, Cloud Scheduling, Resource Management

I. INTRODUCTION

With the rapid growth of cloud computing and the emergence of Quantum Machine Learning (QML), efficient workload prediction and categorization have become critical challenges. Cloud platforms increasingly support heterogeneous workloads, including classical machine learning, deep learning, and emerging quantum-inspired workloads. Accurate prediction and classification of these workloads are essential for efficient resource allocation, scheduling, cost optimization, and performance improvement. Several researchers have proposed AI- and ML-based approaches to optimize cloud workloads, scheduling policies, and resource utilization. This section reviews existing literature related to workload optimization, prediction, and characterization in cloud environments, which form the foundation for extending these techniques toward QML workload prediction and categorization.

II. LITERATURE SURVEY

Priyadarshini et al. [1] presented an AI/ML-based workload optimization framework aimed at enhancing security and scalability in cloud environments. Their approach improved resource utilization and system efficiency by analyzing workload behavior patterns. However, the study focused primarily on classical AI/ML workloads and did not consider the unique characteristics of quantum or hybrid quantum-classical workloads.

Liu et al. [2] proposed Ares, a fair and efficient scheduling mechanism for deep learning jobs using elastic fair queuing. The scheduler improved fairness and reduced job starvation in shared cloud clusters. Although effective for deep learning workloads, the approach lacks workload-type prediction capabilities and does not address quantum machine learning workloads with distinct execution and resource requirements.

Sampling-based multi-job placement for heterogeneous deep learning clusters was introduced by Liu et al. [3]. Their method improved cluster throughput by efficiently placing multiple jobs across heterogeneous resources. While the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, January 2026 Impact Factor: 7.67

approach handles workload diversity, it assumes classical GPU-based training jobs and does not support workload categorization for quantum or hybrid computational models.

Liu et al. [4] proposed SMore, a serverless-based co-location scheduling strategy to enhance GPU utilization in deep learning clusters. The model achieved higher resource efficiency by intelligently co-locating workloads. However, the study is limited to deep learning workloads and does not include workload prediction mechanisms for emerging QML applications that may require different execution models.

Lin et al. [5] developed a universal performance modeling framework for machine learning training on multi-GPU platforms. Their model accurately predicted performance across various configurations. Despite its effectiveness, the framework is designed for classical ML training and does not incorporate workload classification or prediction for quantum machine learning tasks.

UniSched, proposed by Gao et al. [6], introduced a unified scheduler capable of handling deep learning training jobs with varying user demands. The scheduler improved overall system performance by considering workload priorities. Nevertheless, the study does not address workload categorization or predictive modeling for quantum or hybrid workloads in cloud environments.

Hu et al. [7] presented a workload characterization methodology using supervised and unsupervised deep learning techniques. Their approach effectively classified workloads based on execution behavior and resource usage patterns. This work is highly relevant to workload categorization; however, it is limited to classical workloads and does not explore quantum workload characteristics.

Li et al. [8] proposed an interference-aware opportunistic job placement technique for shared distributed deep learning clusters. Their method reduced performance degradation caused by workload interference. Although effective, the approach does not incorporate workload prediction models and does not consider the distinct execution patterns of QML workloads.

III. ANALYSIS OF EXISTING WORKS

TABLE I: Analysis Of Existing Works

Ref	Author	Year	Focus Area	Methodology	Key Contribution	Limitation
1	Priyadarshini et al.	2024	Cloud workload optimization	AI/ML-based analysis	Improved scalability and security	No QML support
2	Liu et al.	2025	DL job scheduling	Elastic fair queuing	Fair scheduling	No workload prediction
3	Liu et al.	2024	Job placement	Sampling-based placement	Improved throughput	Classical workloads only
4	Liu et al.	2025	GPU utilization	Serverless co-location	Higher GPU efficiency	No QML consideration
5	Lin et al.	2024	Performance modelling	Universal ML model	Accurate performance prediction	No workload categorization
6	Gao et al.	2024	Unified scheduling	Demand-aware scheduling	Better system utilization	No predictive analysis
7	Hu et al.	2024	Workload characterization	Deep learning models	Effective classification	Classical workloads only
8	Li et al.	2024	Job placement	Interference-aware placement	Reduced interference	No quantum workload study

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, January 2026
Impact Factor: 7.67

IV. MAJOR CHALLENGES

Despite the promising potential of Quantum Machine Learning (QML) for enhancing cloud intelligence, several critical challenges hinder its practical adoption for workload prediction and categorization in cloud environments. These challenges arise from limitations in quantum hardware, integration complexities with classical cloud systems, and the inherent uncertainty of quantum computations.

a. Limited Availability of Scalable Quantum Hardware: Current quantum processors operate in the Noisy Intermediate-Scale Quantum (NISQ) era, offering limited qubit counts, high error rates, and short coherence times. These constraints restrict the size and complexity of QML models that can be deployed for real-time workload prediction in large-scale cloud platforms. As a result, many proposed QML-based workload models remain theoretical or are validated only through simulation.

b. Hybrid Integration Complexity : Cloud infrastructures are predominantly classical, whereas QML workloads require hybrid quantum–classical execution pipelines. Designing seamless coordination between classical cloud schedulers and quantum accelerators introduces architectural complexity, increased latency, and synchronization overhead. Efficient orchestration mechanisms for hybrid workflows remain an open research problem.

c. Quantum Data Encoding Overhead : Mapping classical cloud workload metrics (CPU usage, memory demand, I/O patterns) into quantum states requires specialized encoding techniques such as amplitude encoding or angle encoding. These encoding processes can be computationally expensive and may negate the theoretical speedup offered by quantum models, particularly for high-dimensional workload datasets.

d. Lack of Standardized QML Benchmarks :Unlike classical ML workloads, there is no widely accepted benchmark dataset or evaluation framework for QML-based workload prediction. The absence of standardized metrics, datasets, and experimental protocols makes it difficult to compare different QML approaches objectively and assess their effectiveness in real cloud environments.

e. Noise and Prediction Reliability :Quantum computations are inherently probabilistic and sensitive to noise. This uncertainty can lead to unstable predictions, which is problematic for cloud resource management tasks that require high reliability and deterministic behavior. Ensuring consistent prediction accuracy under noisy quantum execution remains a significant challenge.

f. High Cost and Limited Accessibility :Access to real quantum hardware is costly and often restricted through cloud-based quantum services. This limits large-scale experimentation and slows down the validation of QML workload prediction models. Additionally, pricing models for quantum resources are not yet optimized for continuous cloud workload analysis.

g. Skill and Toolchain Gap :The development of QML-based cloud solutions requires expertise in quantum computing, machine learning, and cloud systems. The lack of mature development tools, debugging frameworks, and skilled professionals poses a barrier to widespread adoption and industrial deployment.

h. Security and Privacy Concerns :Integrating quantum components into cloud systems introduces new attack surfaces. Secure transmission of workload data to quantum processors, protection of quantum circuits, and privacy preservation in hybrid execution environments require further investigation, particularly for multi-tenant cloud platforms.

V. RESEARCH GAP IDENTIFICATION

From the reviewed literature, it is evident that significant progress has been made in optimizing, scheduling, and characterizing classical machine learning and deep learning workloads in cloud environments. However, most existing approaches do not address workload prediction and categorization for Quantum Machine Learning workloads. QML workloads exhibit distinct execution patterns, hybrid quantum-classical computation models, and unique resource requirements that are not captured by traditional workload models. Additionally, current studies lack predictive frameworks capable of identifying and classifying QML workloads before execution. These limitations highlight the need for a dedicated workload prediction and categorization approach tailored for QML workloads in cloud computing environments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, January 2026
Impact Factor: 7.67

VI. PROPOSED METHODOLOGY

A. Data Collection and Preprocessing

The methodology begins with collecting workload traces from cloud-based ML platforms and quantum simulators. Both classical ML workloads and quantum or quantum-inspired workloads are included to ensure diversity. Preprocessing techniques such as normalization, noise removal, and feature selection are applied to prepare the dataset for modeling.

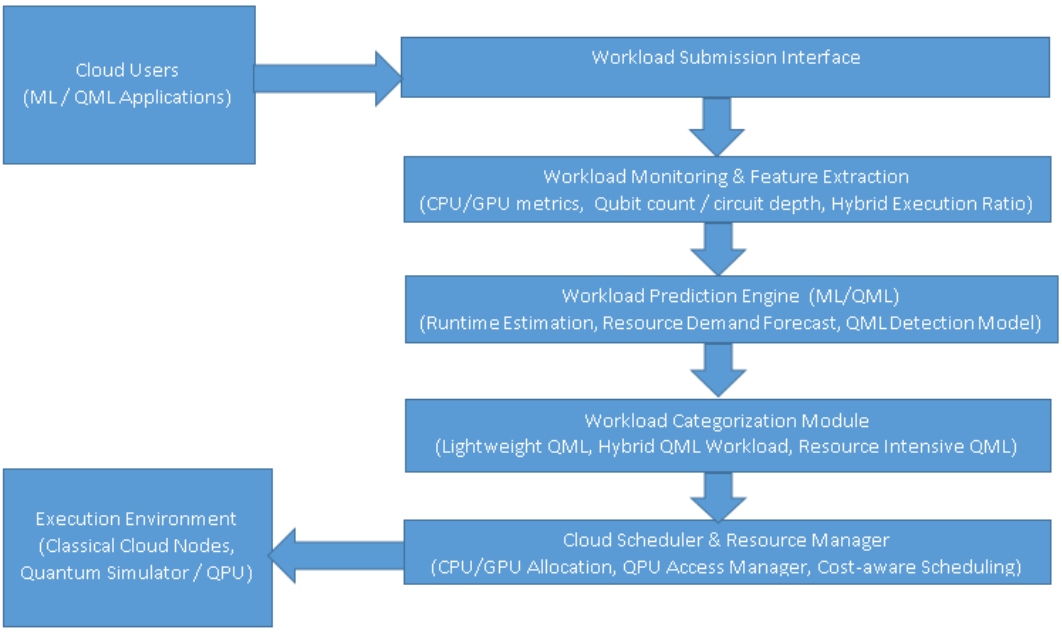

Fig. 1 A Sample flow diagram of proposed methodology

B. Feature Engineering

Key workload features are derived, including:

- Number of qubits and quantum circuit depth
- Hybrid execution ratio (quantum vs classical)
- Resource usage patterns (CPU, GPU, memory)
- Execution latency and job duration
- Historical scheduling behavior

These features help differentiate QML workloads from traditional ML workloads.

C. Workload Prediction Model

A machine learning-based prediction model is employed to identify workload characteristics prior to execution. Supervised learning techniques such as Random Forest or Gradient Boosting can be used for prediction, while unsupervised models such as clustering help detect unknown workload patterns. The prediction model estimates execution cost, resource demand, and workload type.

D. Workload Categorization Framework

Based on prediction outcomes, workloads are categorized into multiple classes:

- Lightweight QML workloads
- Hybrid quantum–classical workloads
- Resource-intensive quantum simulation workloads

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, January 2026 Impact Factor: 7.67

This categorization enables the cloud scheduler to apply workload-specific resource allocation and scheduling strategies.

E. Performance Evaluation

The proposed framework is evaluated using metrics such as prediction accuracy, categorization efficiency, resource utilization, execution latency, and scheduling overhead. Comparative analysis with existing cloud workload management techniques demonstrates the effectiveness of the proposed approach.

VII. FUTURE RESEARCH DIRECTIONS

Although existing research has significantly advanced cloud workload optimization and scheduling for classical machine learning applications, several open research challenges remain in the context of Quantum Machine Learning (QML). Future research can focus on developing hybrid workload prediction models that combine classical machine learning techniques with quantum-inspired features to accurately identify QML workloads in cloud environments. Additionally, there is a need for dynamic workload categorization frameworks capable of adapting to evolving quantum hardware constraints and execution patterns. Integrating cost-awareness, energy efficiency, and execution latency into QML workload prediction models represents another promising research direction. Furthermore, validating such models on real-world hybrid quantum–classical cloud platforms will be essential to ensure practical applicability and scalability.

VIII. CONCLUSION

This study reviewed cloud workload optimization and characterization techniques with emphasis on Quantum Machine Learning (QML) workloads. It highlighted that existing cloud solutions are largely tailored for classical ML/DL and fail to capture the unique execution and resource patterns of QML applications. To address this gap, a QML-focused workload prediction and categorization methodology was outlined, enabling proactive scheduling and efficient resource utilization. As quantum–classical integration grows in cloud platforms, such workload-aware strategies will be critical for scalable and efficient hybrid computing, providing a foundation for future intelligent cloud management research.

IX. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all researchers and scholars whose prior work in cloud computing, workload optimization, and Quantum Machine Learning has provided the foundation for this study. Special thanks are extended to the academic mentors and reviewers for their valuable guidance, constructive feedback, and continuous encouragement throughout the research process. The authors also acknowledge the institutional support and computational resources that enabled the successful completion of this work. Finally, appreciation is extended to colleagues and peers for their insightful discussions and support during the preparation of this paper.

REFERENCES

[1]. Priyadarshini, S., Sawant, T.N., Bhimrao Yadav, G., Premalatha, J. and Pawar, S.R., “Enhancing security and scalability by AI/ML workload optimization in the cloud”, Cluster Computing, vol. 27, no. 10, pp.13455-13469, 2024.
[2]. Liu, Y., Chen, C., Wang, Q., Feng, Y., Cui, W., Chen, Q. and Guo, M., “Ares: Fair and Efficient Scheduling of Deep Learning Jobs with Elastic Fair Queuing”, ACM Transactions on Architecture and Code Optimization, 2025.
[3]. Liu, K., Wang, J., Huang, Z. and Pan, J., “Sampling-Based Multi-Job Placement for Heterogeneous Deep Learning Clusters”, IEEE Transactions on Parallel and Distributed Systems, vol. 35, no. 6, pp.1029-1043, 2024.
[4]. Liu, J., Cai, Z., Liu, Y., Li, H., Zhang, Z., Ma, R. and Buyya, R., “SMore: Enhancing GPU Utilization in Deep Learning Clusters by Serverless-Based Co-Location Scheduling”, IEEE Transactions on Parallel and Distributed Systems, 2025.
[5]. Lin, Z., Sun, N., Bhattacharya, P., Feng, X., Feng, L. and Owens, J.D., “Towards Universal Performance Modeling for Machine Learning Training on Multi-GPU Platforms”, IEEE Transactions on Parallel and Distributed Systems, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, January 2026 Impact Factor: 7.67

[6]. Gao, W., Ye, Z., Sun, P., Zhang, T. and Wen, Y., "UniSched: A unified scheduler for deep learning training jobs with different user demands", IEEE Transactions on Computers, vol. 73, no. 6, pp.1500-1515, 2024.
[7]. Hu, B., Kempf, K. and Mason, N., "A Workload Characterization Methodology Using Supervised and Unsupervised Deep Learning", IEEE Access, 2024.
[8]. Li, H., Zhao, H., Sun, T., Li, X., Xu, H. and Li, K., "Interference-aware opportunistic job placement for shared distributed deep learning clusters", Journal of Parallel and Distributed Computing, vol. 183, pp.104776, 2024.
[9]. Sharma, C., Sharma, S., Kautish, S., Alsallami, S.A., Khalil, E.M. and Mohamed, A.W., "A new median-average round Robin scheduling algorithm: An optimal approach for reducing turnaround and waiting time", Alexandria Engineering Journal, vol. 61, no. 12, pp.10527-10538, 2022.
[10]. Wang, X., Wang, X. and Zhang, S., "Adverse drug reaction detection from social media based on quantum bi-lstm with attention", IEEE Access, vol. 11, pp.16194-16202, 2022.
[11]. Ghasemi, M., Akbari, M.A., Zare, M., Mirjalili, S., Deriche, M., Abualigah, L. and Khodadadi, N., "Birds of prey-based optimization (BPBO): a metaheuristic algorithm for optimization", Evolutionary Intelligence, vol. 18, no. 4, pp.1-68, 2025.
[12]. Josphineleela, R., Kumar, G.S., Ramesh, T. and Balamurugan, K.S., "Optimized multiple objects tracking with conformalized graph neural network and narwhal optimizer for embedded system IoT and mobile edge computing", Ain Shams Engineering Journal, vol. 16, no. 10, pp.103581, 2025.
[13]. Mahapatra, A., Pradhan, R., Majhi, S.K. and Mishra, K., "Quantum ml-based cooperative task orchestration in dew-assisted IoT framework", Arabian Journal for Science and Engineering, vol. 50, no. 15, pp.11975-12002, 2025.
[14]. Rashid, A. and Chaturvedi, A., "Cloud computing characteristics and services: a brief review", International Journal of Computer Sciences and Engineering, vol. 7, no. 2, pp.421-426, 2019.
[15]. Saxena, D., Kumar, J., Singh, A.K. and Schmid, S., "Performance analysis of machine learning centered workload prediction models for cloud", IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp.1313-1330, 2023.
[16]. González-Martínez, J.A., Bote-Lorenzo, M.L., Gómez-Sánchez, E. and Cano-Parra, R., "Cloud computing and education: A state-of-the-art survey", Computers & Education, vol. 80, pp.132-151, 2015.
[17]. De Simone, V., Di Pasquale, V., Calabrese, J., Miranda, S. and Iannone, R., "A Supervised Machine Learning-Based Approach for Task Workload Prediction in Manufacturing: A Case Study Application", Machines, vol. 13, no. 7, pp.602, 2025.
[18]. Yu, Y., Jindal, V., Bastani, F., Li, F. and Yen, I.L., "Improving the smartness of cloud management via machine learning based workload prediction", In Proceedings of IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Vol. 2, pp. 38-44, July, 2018.
[19]. Kumar, J., Singh, A.K. and Buyya, R., "Self-directed learning based workload forecasting model for cloud resource management", Information Sciences, vol. 543, pp.345-366, 2021.
[20]. Sabyasachi, A.S., Sahoo, B.M. and Ranganath, A., "Deep cnn and lstm approaches for efficient workload prediction in cloud environment", Procedia Computer Science, vol. 235, pp.2651-2661, 2024.
[21]. Chen, Z., Hu, J., Min, G., Zomaya, A.Y. and El-Ghazawi, T., "Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning", IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 4, pp.923-934, 2019.
[22]. Saxena, D., Singh, A.K. and Buyya, R., "OP-MLB: An online VM prediction-based multi-objective load balancing framework for resource management at cloud data center", IEEE Transactions on Cloud Computing, vol. 10, no. 4, pp.2804-2816, 2021.
[23]. Saxena, D., Gupta, I., Kumar, J., Singh, A.K. and Wen, X., "A secure and multiobjective virtual machine placement framework for cloud data center", IEEE Systems Journal, vol. 16, no. 2, pp.3163-3174, 2021.
[24]. Adil, M., Nabi, S., Aleem, M., Diaz, V.G. and Lin, J.C.W., "CA‐MLBS: content‐aware machine learning based load balancing scheduler in the cloud environment", Expert Systems, vol. 40, no. 4, pp.e13150, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, January 2026
Impact Factor: 7.67

[25]. Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y., "Deep learning", Cambridge: MIT press, Vol. 1, No. 2, 2016.

[26]. Shahab, E. and Taghipour, S., "Designing a resilient cloud network fulfilled by quantum machine learning", International Journal of Management Science and Engineering Management, pp.1-11, 2025.