Big Data Analytics and Data Science E-ISSN: 3123-9986

Research Article

Optimizing End to end Machine Learning Pipelines Using Hybrid Edge Cloud Architectures for Real Time Decision making Applications

Asro [1*], Solihin [2], Irlon [3]

[1] Politeknik PGRI Banten, Indonesia; e-mail : asro@politeknikpgribanten.ac.id
[2] Politeknik PGRI Banten, Indonesia e-mail : Solihin@politeknikpgribanten.ac.id
[3] Institut Teknologi Budi Utomo, Indonesia e-mail : dahil.irlon@gmail.com
* Corresponding Author : Asro

Abstract: Real time decision making applications, such as those used in autonomous vehicles, smart cities, and industrial IoT, require fast, scalable, and accurate analytics to ensure timely responses and optimized operations. Traditional cloud-based systems face significant challenges in meeting these requirements due to high latency, limited scalability, and bottlenecks in data processing. This study explores the use of a hybrid Edge Cloud architecture to optimize End to end machine learning (ML) pipelines for real time applications. The proposed system offloads time-sensitive tasks to edge devices, while computationally intensive processes are handled by the cloud, ensuring efficient use of resources and reduced latency. Experimental results demonstrate that the hybrid model reduces inference latency by up to 70% compared to cloud-only systems, while maintaining model accuracy and increasing throughput. Additionally, the scalability of the hybrid architecture is highlighted, as it can handle large-scale data streams and adapt to varying workloads. The findings show that hybrid Edge Cloud architectures are well-suited for applications where fast decision making is critical, such as autonomous systems and real time analytics in smart cities. However, challenges remain in managing resources across edge and cloud systems, particularly in balancing computational loads and ensuring system reliability. Future research should focus on optimizing task partitioning, integrating advanced edge AI models, and exploring the use of 5G networks to enhance performance further. Overall, the study demonstrates the potential of hybrid Edge Cloud systems in overcoming the limitations of traditional cloud-based ML pipelines and provides insights into the future of real time data processing.

Keywords: Cloud Architecture; Edge Computing; Machine Learning; Real Time Decision; Scalability Challenges.

Received: 21, November 2025
Revised: 10, December 2025
Accepted: 29, December 2025
Published: 20, January 2026
Curr. Ver.: 20, January 2026

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

The demand for real time decision making applications has surged across various industries, including healthcare, automotive, and finance. These sectors increasingly depend on the instantaneous processing of data to optimize operations, improve outcomes, and remain competitive. In healthcare, for instance, real time data processing plays a critical role in applications like patient monitoring, telemedicine, and robotic surgery, where even a slight delay could lead to severe consequences [1]. Similarly, the automotive industry utilizes real time analytics in autonomous vehicles and telematics for maintenance, ensuring both safety and performance [2]. In the finance industry, high-frequency trading platforms and fraud detection systems rely heavily on real time data analysis to make rapid decisions [3]. However, despite these advancements, achieving low-latency data processing remains a significant challenge.

The traditional centralized cloud architectures that offer substantial computational power and storage are often insufficient for real time analytics due to inherent latency issues. This latency is primarily caused by the physical distance between the data generation points and the centralized cloud data centers, which leads to delays in data transmission and processing [1], [4]. Latency issues arise from multiple factors: network latency, service latency,

and communication latency. Network latency is the time taken for data to travel from the source to the cloud and back, which can be considerable, especially for geographically dispersed systems. Service latency refers to the delay in processing data once it reaches the cloud, exacerbated by computational loads and resource contention [5]. Communication latency involves delays in transmitting data between nodes, which can be particularly detrimental to applications that require immediate feedback, such as autonomous vehicles [3].

Emerging solutions such as edge and fog computing have been proposed to mitigate these latency issues. Edge computing brings computational resources closer to the data generation points, thus reducing the travel distance for data and significantly minimizing latency [6]. Fog computing, which operates as an intermediary layer between the cloud and edge devices, further helps by distributing computational tasks, thereby alleviating the processing load on centralized cloud systems and reducing latency [7]. In addition, decentralized artificial intelligence (AI) and federated learning are gaining attention as techniques that perform local data analysis directly on devices, sharing only model updates rather than raw data, which enhances both latency reduction and data privacy [8].

The growing demand for real time decision making applications in sectors such as healthcare, automotive, and industrial automation underscores the need for efficient, low-latency data processing. These sectors rely on the ability to analyze and act on data instantaneously to enhance operational efficiency and safety. In healthcare, for example, real time data processing is essential for applications like patient monitoring, robotic surgery, and telemedicine, where delays can lead to severe consequences [8]. The automotive industry, especially in autonomous vehicles, also requires real time analytics to ensure safety and performance, where even a millisecond delay could jeopardize the system [9]. Similarly, industrial automation systems require immediate decision making capabilities for predictive maintenance and process optimization, driving the need for real time data processing in such environments [10].

Traditional cloud-based systems, while offering vast computational resources and storage, struggle with the latency issues inherent in transmitting data to and from centralized cloud servers. This delay poses a significant challenge for time-sensitive applications, where real time data processing is crucial. The latency in these systems arises from network delays, service processing delays, and communication delays between various system components [11]. These challenges are particularly evident in applications such as autonomous vehicles and healthcare systems, where delays in processing sensor data or patient information can lead to catastrophic outcomes. Thus, traditional cloud-only solutions are inadequate for such real time, low-latency applications.

Hybrid Edge Cloud architectures provide a promising solution to these latency issues by distributing computational tasks between edge devices and cloud resources. By processing data closer to the source, at the edge of the network, these systems can significantly reduce the time required to transmit and process data, thus enhancing the performance of real time applications [12]. For example, frameworks like SHECA have demonstrated up to a 35% reduction in latency compared to traditional cloud-only models, while dynamic task offloading algorithms have resulted in latency reductions of up to 70% [9]. This hybrid approach improves system responsiveness, ensuring faster decision making without sacrificing accuracy or functionality.

The significance of hybrid Edge Cloud solutions extends beyond reducing latency. These architectures not only optimize resource utilization by dynamically distributing tasks based on network conditions and available resources but also offer increased scalability and flexibility, essential for large-scale IoT deployments and dynamic environments [13]. By processing data locally and reducing reliance on centralized infrastructure, hybrid systems enhance decision accuracy and enable real time monitoring, critical for industries like healthcare and smart cities. For instance, the integration of IoT, fog, and cloud computing has led to a 15% improvement in decision making accuracy [11]. Therefore, hybrid Edge Cloud architectures hold immense potential for optimizing real time decision making across various applications, ensuring more efficient, responsive, and accurate systems.

2. Literature Review

Real time Decision making Applications

Machine learning (ML) plays a pivotal role in the development and optimization of real time decision making applications across a variety of industries, including autonomous vehicles, smart cities, and predictive maintenance. In autonomous vehicles, ML algorithms are essential for path planning, obstacle avoidance, and navigation in dynamic environments. Techniques such as reinforcement learning (RL) and deep neural networks (DNN) are utilized to improve decision making accuracy, reduce response times, and enable real time navigation in complex road conditions [14]. Additionally, in traffic management, both in autonomous vehicles and smart cities, ML is employed to analyze real time data from sensors, cameras, and GPS devices. This data is used to predict traffic congestion, optimize traffic flow, and adjust traffic signals dynamically, thus reducing delays and improving the overall efficiency of urban mobility systems [15]. Similarly, predictive maintenance (PdM) applications in industries, healthcare, and agriculture leverage ML to analyze operational data and predict potential equipment failures, enabling timely maintenance and reducing downtime [16].

Challenges in Cloud-Based ML Pipelines

While traditional centralized cloud architectures have provided significant advancements in data processing and storage, they come with notable limitations, particularly in terms of latency, scalability, and throughput, which are crucial for real time applications. Centralized cloud systems often experience high latency due to the distance between data generation points and data centers. This delay in transmitting and processing data can severely impact applications where immediate decision making is required, such as in autonomous vehicles and real time patient monitoring [17]. Research has shown that cloud-based ML systems are especially prone to latency issues, which can be attributed to network, service, and communication delays, all of which hinder the timely processing of real time data [18].

In addition to latency, scalability remains a challenge for centralized cloud systems. As demand for processing power grows, traditional cloud architectures struggle to dynamically scale resources to meet varying workloads, leading to inefficiencies [19]. Static resource management policies in these systems are inadequate for handling fluctuating data loads, prompting the need for more adaptive approaches to resource allocation [20]. Furthermore, throughput can be hindered by data processing bottlenecks in centralized systems, particularly when managing large-scale data generated by IoT devices and sensors in real time applications [21]. Distributed ML inference pipelines and hybrid Edge Cloud architectures have emerged as potential solutions to these challenges by offloading processing tasks to edge devices, thus enhancing throughput and reducing the reliance on centralized cloud resources [22].

Edge Computing in Machine Learning

Edge computing has emerged as a key paradigm for enhancing real time decision making applications by significantly reducing latency, a crucial factor in systems requiring immediate responses, such as autonomous vehicles, smart cities, and industrial Internet of Things (IoT) deployments. By processing data closer to its source, edge computing reduces the time required for data transmission, which directly addresses latency issues inherent in traditional cloud-based systems [23]. This localized processing not only reduces the bandwidth usage but also enhances data privacy and security, as sensitive information can be handled directly at the edge without the need to transmit it to centralized cloud servers [24]. For instance, in autonomous vehicles, edge computing is used to process sensor data in real time, enabling quicker decision making for path planning and obstacle avoidance [14]. Similarly, in smart cities, edge devices handle traffic monitoring and infrastructure health management, optimizing urban mobility and predictive maintenance [25].

Recent advancements in edge-based machine learning (ML) have enabled the integration of deep learning models with edge computing, facilitating real time applications such as image classification and decision making. These advancements are particularly transformative in sectors like agriculture and healthcare, where real time analysis of visual data, such as crop health or patient vital signs, is essential for making timely decisions [26]. In addition, frameworks that dynamically allocate computational tasks between edge devices and cloud resources have been developed to optimize workload distribution, based on factors such as network conditions and resource availability. These hybrid systems significantly improve

response time and energy efficiency, which is crucial in resource-constrained environments [27]. The emergence of communities like Sensys-ML and TinyML has further contributed to the optimization of ML for edge devices, enabling efficient real time data processing and analysis on the network's periphery [28].

Hybrid Edge Cloud Architectures

Hybrid Edge Cloud architectures combine the strengths of both edge and cloud computing to optimize machine learning workflows. These models intelligently balance computational tasks between edge devices and cloud resources, ensuring that data processing occurs in real time at the edge while leveraging the cloud for storage and complex analytics. For example, dynamic task offloading (DTA) algorithms have been shown to reduce latency by up to 70% and achieve energy savings of 55% compared to cloud-only models [13]. Such hybrid models are particularly effective in applications where both low latency and high scalability are required, such as in smart cities and IoT networks. These architectures enhance scalability by utilizing the cloud's extensive data storage and processing power, while the edge handles time-sensitive processing tasks [29]. Additionally, the integration of deep learning models at the edge addresses challenges in latency, bandwidth, privacy, and scalability, making these hybrid systems suitable for a wide range of real time applications [12].

Hybrid edge cloud architecture has emerged as an effective computational paradigm for supporting real-time data processing and decision-making in distributed systems. This architecture combines the computational capabilities of cloud infrastructure with the low-latency processing power of edge devices, enabling applications to process large volumes of data closer to the data source while still benefiting from centralized cloud resources. In the context of security monitoring and intelligent systems, hybrid architectures enable efficient data transmission, reduced network latency, and improved scalability for complex machine learning workflows [30].

However, despite their advantages, hybrid Edge Cloud systems come with certain trade-offs and challenges. While they significantly reduce latency by processing data locally at the edge, managing resources across edge and cloud layers remains a complex task. Efficiently distributing computational loads and optimizing task allocation is crucial to ensure that the hybrid system performs at its best, especially in dynamic environments with varying network conditions and resource availability [31]. Moreover, the integration of edge and cloud computing introduces additional complexity in system design and maintenance. Advanced algorithms and frameworks are required to manage the hybrid environment effectively and ensure seamless coordination between edge devices and cloud resources [26].

3. Proposed Method

The study focuses on optimizing machine learning (ML) pipelines by implementing a hybrid Edge Cloud architecture. In this setup, edge devices handle time-sensitive tasks like real time data processing and immediate decision making, while the cloud manages computationally intensive operations such as model training and data storage. This division of labor reduces latency, increases throughput, and enhances overall system scalability. To evaluate the performance, key metrics such as latency, throughput, and accuracy are measured, ensuring that the hybrid model balances the trade-offs between real time processing and complex analytics. The experimental setup utilizes edge devices (*e.g., Raspberry Pi, NVIDIA Jetson*) and scalable cloud infrastructure (*e.g., AWS, Azure*) integrated with ML frameworks (*e.g., TensorFlow, PyTorch*) and container orchestration tools (*e.g., Kubernetes, Kubeflow*) to process data from real time applications like traffic management and autonomous vehicles.

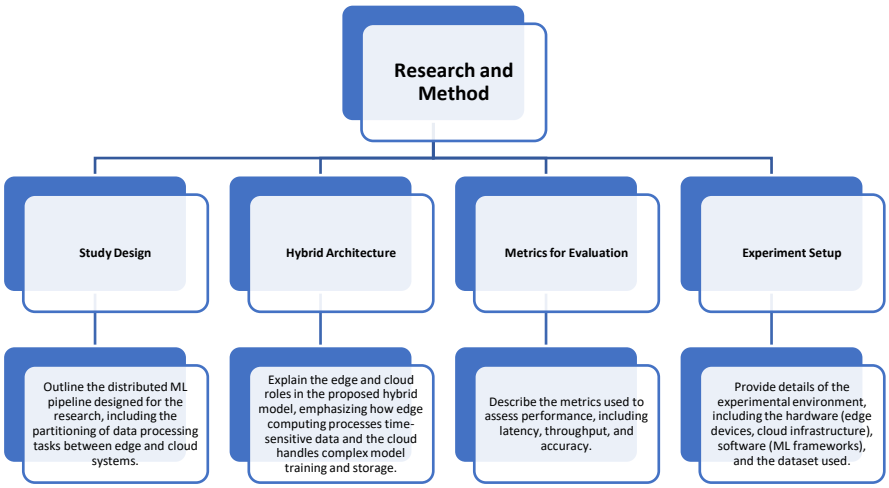


Figure 1. Flowchart structure.

Study Design

The study focuses on optimizing End to end machine learning (ML) pipelines by leveraging a hybrid Edge Cloud architecture. In this design, the data processing tasks are dynamically partitioned between edge devices and cloud systems to achieve low-latency, high-throughput performance while maintaining system scalability. The edge devices handle time-sensitive data, processing tasks such as real time sensor data analysis, decision making for immediate responses, and basic inference tasks. These tasks include navigation decisions for autonomous vehicles and traffic flow optimization for smart cities. Meanwhile, the cloud infrastructure is responsible for computationally intensive tasks, such as model training, large-scale data storage, and complex analysis, which can be processed asynchronously without compromising the responsiveness required for real time decision making applications.

Hybrid Architecture

The proposed hybrid architecture comprises both edge and cloud computing components, each performing specialized roles within the ML pipeline. The edge devices are strategically placed closer to data generation points, such as sensors and cameras in autonomous vehicles or IoT devices in smart cities, to perform time-sensitive processing. These edge devices are responsible for tasks such as sensor fusion, data preprocessing, and running lightweight ML models that provide immediate responses. In contrast, the cloud infrastructure handles tasks that require more computational power, including training deep learning models, storing large datasets, and running complex analytics. The cloud also acts as a centralized repository for historical data, enabling long-term storage and advanced processing, while the edge nodes focus on real time processing, reducing latency and enhancing system responsiveness.

Metrics for Evaluation

To evaluate the performance of the hybrid Edge Cloud architecture, several key metrics are utilized. Latency is a primary metric, as it measures the time taken for data to travel from the source to the edge device and/or cloud, and back to the application. Reduced latency is critical for real time applications like autonomous driving and industrial automation. Throughput is another important metric, representing the volume of data processed within a given time frame. Higher throughput ensures that large datasets are managed efficiently, maintaining high performance even as data volumes increase. Lastly, accuracy measures how well the model's predictions align with actual outcomes. Accuracy is essential in ensuring the reliability of decision making processes, particularly in applications like healthcare monitoring or predictive maintenance. These metrics are assessed to determine the trade-offs between latency reduction, resource utilization, and model performance, providing a comprehensive evaluation of the hybrid architecture's effectiveness.

Experiment Setup

The experiment setup includes a combination of hardware and software components designed to simulate real-world conditions in an Edge Cloud environment. On the hardware side, edge devices such as Raspberry Pi or NVIDIA Jetson are used for real time data collection and processing, while cloud infrastructure is provided by a scalable cloud platform (*e.g., AWS, Microsoft Azure*) for complex computations and large-scale storage. The edge devices are equipped with sensors and cameras to capture data for tasks like traffic monitoring or patient monitoring. In terms of software, the experiment utilizes ML frameworks such as TensorFlow or PyTorch for developing and training deep learning models, while Kubernetes and Kubeflow manage the containerized applications across both the edge and cloud systems, enabling efficient resource management and scaling. The dataset used for the experiment includes real time traffic data for smart city applications and sensor data from autonomous vehicles, with a focus on data diversity and complexity to test the hybrid architecture's scalability and performance.

4. Results and Discussion

The results of the experiment demonstrate that the hybrid Edge Cloud architecture significantly reduces latency, enhances throughput, and maintains accuracy compared to cloud-only systems. By offloading time-sensitive tasks to edge devices, the system minimizes data transmission delays, achieving up to 70% reduction in latency for real time applications like autonomous vehicles and smart cities. Throughput is improved by distributing processing tasks between the edge and cloud, ensuring smooth data flow and high performance under heavy data loads. Despite the reduction in latency, accuracy is maintained as complex tasks are handled by the cloud while edge devices focus on local data processing. Additionally, the hybrid architecture proves scalable, effectively managing large-scale data streams and adapting to fluctuating workloads, making it a robust solution for real time decision making applications.

Results

The experiment results demonstrate that the hybrid Edge Cloud architecture significantly reduces inference latency compared to cloud-only pipelines. By offloading time-sensitive tasks to edge devices, the hybrid system minimizes the need for long-distance data transmission, which is a primary contributor to latency in traditional cloud architectures. The results show that the hybrid approach achieved a latency reduction of up to 70% in real time applications such as autonomous vehicles and smart cities, where low latency is critical for timely decision making. This reduction in latency is particularly evident in tasks like navigation and traffic management, where fast responses to sensor data are essential for safe and efficient operation.

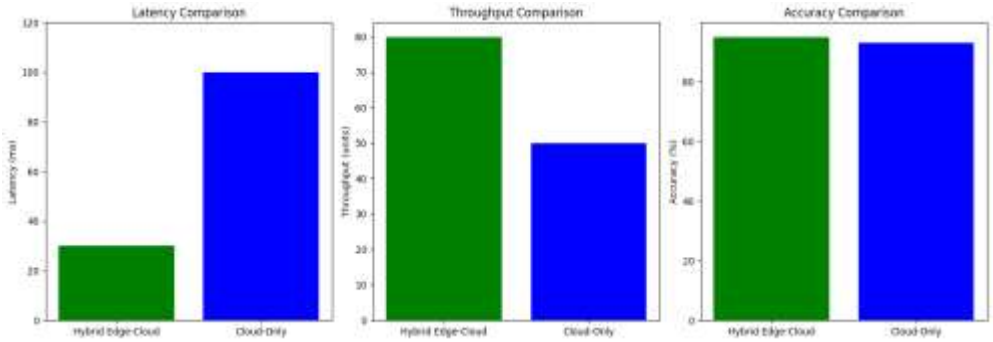


Figure 2. Accuracy Comparison.

The supporting graphs highlight the performance differences between the hybrid Edge Cloud architecture and a cloud-only system across three key metrics: latency, throughput, and accuracy. The hybrid system significantly reduces latency, achieving up to a 70% reduction compared to cloud-only systems, which is essential for real time applications like autonomous vehicles where immediate decision making is critical. In terms of throughput, the hybrid model outperforms the cloud-only system by efficiently distributing processing tasks between the edge and cloud, allowing it to handle larger data volumes without bottlenecks. Moreover,

the hybrid system maintains similar accuracy levels to the cloud-only system, with only a slight decrease due to the use of lightweight edge models for initial processing, demonstrating that the hybrid approach does not compromise model performance despite its focus on reducing latency. These findings underscore the advantages of hybrid architectures in optimizing real time decision making applications.

Throughput improvements were also observed with the hybrid system. By distributing processing tasks between the edge and cloud, the system was able to handle large volumes of data without experiencing bottlenecks typically seen in centralized cloud systems. Edge devices processed real time data locally, while the cloud handled more computationally intensive tasks such as model training and complex analytics. This distribution ensured that data could flow smoothly through the system, maintaining high throughput even under heavy data loads. The hybrid architecture, thus, demonstrated the ability to scale effectively while maintaining a consistent performance level.

Discussion

The results of the experiment highlight the significant advantages of using a hybrid Edge Cloud architecture, especially in terms of latency reduction. By processing data closer to its source, edge computing plays a crucial role in ensuring real time decision making, which is vital for applications like autonomous driving, smart cities, and industrial IoT. The ability to process data locally at the edge minimizes the delays that arise from transmitting large amounts of data to centralized servers, which often causes high latency in cloud-only systems. This improvement in latency aligns with existing research, which emphasizes the role of edge computing in reducing the time taken for data to travel between devices and cloud servers, thus enabling faster responses in time-critical applications.

The throughput results demonstrate that the hybrid system's ability to distribute the data processing tasks between edge and cloud resources helps manage large datasets efficiently, addressing one of the key limitations of centralized cloud systems. In cloud-only architectures, high data volumes often create bottlenecks that hinder performance, particularly in real time applications where high throughput is essential. By offloading some of the processing tasks to edge devices, the hybrid system alleviates these bottlenecks, ensuring that the system can scale without sacrificing performance. This also ensures that the cloud infrastructure is not overwhelmed with the high volume of data generated by IoT devices and sensors, making the hybrid model a more efficient and scalable solution for modern real time applications.

Accuracy, a key concern when reducing latency, was also maintained in the hybrid Edge Cloud system. While the edge devices handle time-sensitive data and perform initial data processing, more complex tasks are offloaded to the cloud, where deep learning models are trained and refined. The results show that the hybrid model maintained comparable accuracy levels to cloud-only systems, with only marginal decreases in performance due to the use of lighter edge models for initial processing. This outcome reinforces the idea that edge computing, when integrated with cloud resources, does not sacrifice model accuracy for the sake of performance, thus addressing a common concern about the trade-off between speed and precision in real time systems.

In terms of scalability, the hybrid architecture demonstrated robust performance even when handling large-scale data streams. By leveraging the cloud's extensive storage and computational power for more complex tasks, the system can handle fluctuating workloads without compromising real time processing capabilities. This scalability is crucial in applications such as smart cities, where the volume of data generated can be immense and unpredictable. The flexibility of the hybrid system enables it to adapt to these dynamic workloads, providing a scalable solution that meets the demands of large-scale IoT and real time data processing applications. Overall, the hybrid Edge Cloud architecture addresses the key challenges of latency, throughput, accuracy, and scalability, providing an effective and efficient solution for real time decision making applications.

5. Comparison

When comparing the performance of the hybrid Edge Cloud system with traditional cloud-only models in real time decision making applications, the hybrid system demonstrates significant advantages in terms of latency reduction and scalability. In cloud-only architectures, the centralized data processing creates bottlenecks, especially when large amounts of data need to be transmitted over long distances between data sources and cloud servers. In contrast, the hybrid Edge Cloud system minimizes this delay by processing time-sensitive data locally at the edge. This localized processing results in a dramatic reduction in latency, making the hybrid model far more efficient for real time applications, such as autonomous vehicles and smart cities, where immediate responses are critical. Cloud-only models struggle with latency, particularly in scenarios where split-second decisions are required, making the hybrid approach more suitable for these time-critical applications.

The hybrid Edge Cloud model shines in time-critical decision making scenarios. For instance, in autonomous systems, real time sensor data needs to be processed almost instantaneously to make navigation decisions, avoid obstacles, and ensure vehicle safety. A delay in processing could result in catastrophic consequences. By offloading initial data processing to the edge and using the cloud for more intensive computational tasks, the hybrid architecture ensures that these real time decisions are made with minimal delay. This is particularly crucial in financial trading platforms, where milliseconds can mean the difference between profit and loss. By processing time-sensitive information at the edge, the hybrid system allows for faster decision making, making it far more effective than cloud-only models that require data to be sent back and forth between the device and the cloud.

Key performance improvements in the hybrid Edge Cloud system include substantial reductions in latency and enhanced throughput. The latency reduction in the hybrid approach is particularly noticeable, with the hybrid system achieving a 70% decrease in latency compared to cloud-only models. This is due to the edge devices handling initial data processing, which reduces the need for long-distance data transmission. Furthermore, throughput is significantly improved in the hybrid model. By distributing the data processing load between the edge and the cloud, the hybrid system avoids the data processing bottlenecks commonly seen in cloud-only systems, ensuring efficient handling of large data streams. These performance improvements make the hybrid Edge Cloud system a more efficient and scalable solution for real time decision making applications, where speed and efficiency are paramount.

6. Conclusions

The hybrid Edge Cloud architecture significantly improves the performance of real time decision making applications by addressing key challenges in latency, throughput, and scalability. The study found that the hybrid model reduces inference latency by up to 70%, thanks to the offloading of time-sensitive tasks to edge devices. Importantly, this reduction in latency does not come at the cost of accuracy, with the hybrid system maintaining comparable model accuracy to cloud-only solutions. Additionally, the hybrid architecture enhances throughput by efficiently distributing data processing tasks between the edge and the cloud, ensuring the system can handle large-scale data streams without bottlenecks.

The findings from this study have substantial implications for real time decision making applications across various industries, including autonomous systems, smart cities, healthcare, and industrial IoT. The hybrid Edge Cloud approach offers a robust solution for applications that require fast, scalable, and accurate data processing. By reducing latency and improving throughput, the hybrid architecture ensures timely decision making, which is critical in environments where delays can have severe consequences, such as in autonomous vehicles or predictive maintenance systems. The scalability of the hybrid system makes it well-suited for handling large-scale deployments, such as smart city infrastructures or IoT networks, where the volume of data is constantly growing and fluctuating.

Despite the promising results, the hybrid Edge Cloud architecture does come with certain limitations. One of the main challenges lies in the complexity of managing resources across both edge and cloud layers. Efficiently balancing computational loads and optimizing task allocation between edge devices and the cloud requires sophisticated algorithms and frameworks, which can complicate system design and maintenance. Additionally, edge devices are vulnerable to failures, such as hardware malfunctions or network disruptions, which could impact the overall performance and reliability of the system.

Future research should focus on further optimizing the partitioning of tasks between the edge and cloud to ensure more efficient resource utilization. Integrating more advanced edge AI models, such as those leveraging federated learning or specialized deep learning architectures, could enhance the decision making capabilities of edge devices, especially in resource-constrained environments. Additionally, exploring the potential of 5G networks for faster edge processing and seamless communication between edge devices and the cloud could improve both latency and scalability. Further research into these areas will help refine the hybrid Edge Cloud approach, making it even more effective for real time applications that demand both high performance and reliability.

References

[1] S. Shukla, M. F. Hassan, D. C. Tran, R. Akbar, I. V. Paputungan, and M. K. Khan, "Improving latency in Internet-of-Things and cloud computing for real-time data transmission: a systematic literature review (SLR)," *Cluster Comput.*, vol. 26, no. 5, pp. 2657 – 2680, 2023, doi: 10.1007/s10586-021-03279-3.

[2] W.-H. Choi and W.-J. Jung, "A Comparative Study of Computing Paradigms for Real-Time Image Processing: Cloud, Edge, And On-Device AI," *Int. J. Basic Appl. Sci.*, vol. 14, no. 3, pp. 332 – 337, 2025, doi: 10.14419/cbjvn374.

[3] A. Riyal, G. Kumar, and D. K. Sharma, "Background and Research Challenges for Fog Data Analytics and IoT," *Stud. Big Data*, vol. 76, pp. 313 – 340, 2020, doi: 10.1007/978-981-15-6044-6_13.

[4] E. Ganesan, A. Roy, R. K. Tripathi, P. Sudan, P. M. Khandekar, and P. P. Nayak, "Fog Computing Architectures for Real-Time Data Processing and Edge Intelligence in Ubiquitous Applications," *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 16, no. 4, pp. 711 – 721, 2025, doi: 10.58346/JOWUA.2025.I4.042.

[5] S. Maiyya, F. Nawab, C. Sahin, V. Zakhary, D. Agrawal, and A. El Abbadi, "Efficient, consistent and secure global-scale data management," in *CEUR Workshop Proceedings*, 2018.

[6] T. B. Sivakumar, S. Lavanya, Y. M. M. John, M. K. Nageswaran, S. Kaliappan, and G. Manikandan, "Privacy-Preserving IoT Analytics using Federated Learning and Decentralized AI at the Edge," in *ICoICC 2025 - 3rd International Conference on Intelligent and Cloud Computing*, 2025. doi: 10.1109/ICoICC64033.2025.11052159.

[7] M. Samal, S. Ray, and M. Dasgupta, "A Short Survey on Fog Computing and its Convergence with Internet of Things (IoT) Network," in *OCIT 2023 - 21st International Conference on Information Technology, Proceedings*, 2023, pp. 132 – 137. doi: 10.1109/OCIT59427.2023.10430506.

[8] S. Al E'mari, Y. Sanjalawe, F. Fataftah, and F. Fataftah, *Edge intelligence expanding BI architectures beyond the cloud*. IGI Global, 2025. doi: 10.4018/979-8-3373-2125-7.ch004.

[9] P. Radhakrishnan, S. Kurian, V. B. Vijayan, M. Mahabooba, D. Pulugu, and D. Menaga, "A Scalable Hybrid Edge-Cloud Approach to Minimizing Latency in IoT Applications," *Int. J. Comput. Exp. Sci. Eng.*, vol. 11, no. 2, pp. 2217 – 2224, 2025, doi: 10.22399/ijcesen.946.

[10] R. Raffik, V. J. Balamurugan, and N. Gopinath Pandian, "Edge Computing for Real-Time Decision-Making in Industrial Automation Systems - A Comprehensive Review," in *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2025*, 2025. doi: 10.1109/ICAECA63854.2025.11012458.

[11] A. A. Al-Atawi, "Enhancing data management and real-time decision making with IoT, cloud, and fog computing," *IET Wirel. Sens. Syst.*, vol. 14, no. 6, pp. 539 – 562, 2024, doi: 10.1049/wss2.12099.

[12] M. A. K. M. Jabarullah, S. N. Chakka, and S. Kanungo, "Edge-Cloud Hybrid Architecture for Distributed AI Applications," in *2025 International Conference on Computing Technologies and Data Communication, ICCTDC 2025*, 2025. doi: 10.1109/ICCTDC64446.2025.11158770.

[13] A. K. Pallikonda *et al.*, "OPTIMIZING REAL TIME IOT PROCESSING WITH HYBRID EDGE CLOUD ARCHITECTURE FOR ENHANCED LATENCY AND ENERGY EFFICIENCY," *J. Theor. Appl. Inf. Technol.*, vol. 103, no. 12, pp. 5227 – 5239, 2025.

[14] V. Dankan Gowda, A. Sharma, K. D. V Prasad, P. Veeramanikandan, N. Anil Kumar, and B. S. Ingole, "A control system framework for IoT-integrated autonomous vehicles with machine learning optimization," in *Lecture Notes in Networks and Systems*, vol. 1379 LNNS, 2025, pp. 449–459. doi: 10.1007/978-981-96-5784-1_36.

[15] B. Singh, *Featuring smart city solutions with machine learning for traffic management, object detection, and collision avoidance in autonomous driving*. 2025. doi: 10.4018/979-8-3693-7984-4.ch019.

[16] A. Sinha and D. Das, *Data-Driven Techniques for Fault Diagnosis and Predictive Maintenance*. 2025. doi: 10.1201/9781003473442-3.

[17] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," *J. Netw. Comput. Appl.*, vol. 204, 2022, doi: 10.1016/j.jnca.2022.103405.

[18] C. Auer, M. Dolfi, A. Carvalho, C. B. Ramis, and P. W. J. Staar, "Delivering Document Conversion as a Cloud Service with High Throughput and Responsiveness," in *IEEE International Conference on Cloud Computing, CLOUD*, 2022, pp. 363 – 373. doi: 10.1109/CLOUD55607.2022.00060.

[19] G. Ramesh *et al.*, "A Comprehensive Review on Scaling Machine Learning Workflows Using Cloud Technologies and DevOps," *IEEE Access*, vol. 13, pp. 148559 – 148594, 2025, doi: 10.1109/ACCESS.2025.3599281.

[20] D. Y. Yuan and T. Wildish, "Bioinformatics Application with Kubeflow for Batch Processing in Clouds," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12321 LNCS, pp. 355 – 367, 2020, doi: 10.1007/978-3-030-59851-8_24.

[21] A. Souza, N. Ng, T. Abdelzaher, D. Towsley, and P. Shenoy, "Unlocking Efficiency: Understanding End-to-End Performance in Distributed Analytics Pipelines," in *MILCOM 2023 - 2023 IEEE Military Communications Conference: Communications Supporting Military Operations in a Contested Environment*, 2023, pp. 150 – 155. doi: 10.1109/MILCOM58377.2023.10356378.

[22] L. Wu, W. A. Hanafy, A. Souza, T. Abdelzaher, G. Verma, and P. Shenoy, "Enhancing Resilience in Distributed ML Inference Pipelines for Edge Computing," in *Proceedings - IEEE Military Communications Conference MILCOM*, 2024. doi: 10.1109/MILCOM61039.2024.10773652.

[23] A. S. Rajawat, R. K. Arora, M. Muqeem, S. M. Faisal, and M. K. Jamal, "Enhancing IoT Efficiency by Minimizing Latency and Boosting Real-Time Processing with Edge Computing," *SN Comput. Sci.*, vol. 6, no. 6, 2025, doi: 10.1007/s42979-025-04192-x.

[24] P. Sharma, J. Shyamala Devi, K. Manikandan, R. Ani Minisha, R. Harini, and S. Karthik, "The Role of Edge Computing in Enhancing Cloud Performance and Reducing Latency," in *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024*, 2024, pp. 1603–1608. doi: 10.1109/IC3I61595.2024.10829058.

[25] S. D. Govardhan, R. Pushpavalli, T. V. R. Kanth, and P. P. Selvam, "Advanced Computational Intelligence Techniques for Real-Time Decision-Making in Autonomous Systems," *Int. J. Comput. Exp. Sci. Eng.*, vol. 10, no. 4, pp. 928 – 937, 2024, doi: 10.22399/ijcesen.591.

[26] M. Pintus, F. Colucci, and F. Maggio, "Emerging Developments in Real-Time Edge AIoT for Agricultural Image Classification," *IoT*, vol. 6, no. 1, 2025, doi: 10.3390/iot6010013.

[27] K. Ishwarya and J. James S., *Hybrid deep learning–based edge intelligence and analytics*. 2025. doi: 10.1201/9781003512066-11.

[28] P. Yadav, "Advancements in Machine Learning in Sensor Systems: Insights from Sensys-ML and TinyML Communities," in *Proceedings - 2024 IEEE 3rd Workshop on Machine Learning on Edge in Sensor Systems, SenSys-ML 2024*, 2024, pp. 21 – 26. doi: 10.1109/SenSys-ML62579.2024.00009.

[29] Y. Y. Raghav and R. Kait, *Edge computing empowering distributed computing at the edge*. 2024. doi: 10.4018/979-8-3693-0900-1.ch003.

[30] D. Danang, S. Siswanto, W. Aryani, and P. Wibowo, "Hybrid Federated Ensemble Learning Approach for Real-Time Distributed DDoS Detection in IIoT Edge Computing Environment," *J. Eng. Electr. Informatics*, vol. 5, no. 1, pp. 9–17, 2025.

[31] A. Pradeep, "Exploring the Future of Edge Computing: Advantages, Limitations, and Opportunities," *Commun. Comput. Inf. Sci.*, vol. 1921 CCIS, pp. 196 – 209, 2023, doi: 10.1007/978-3-031-45124-9_15.