International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 1-18

https://doi.org/10.70917/ijcisim-2026-2027
Article

A Semantic-Aware Data Offloading Approach for Latency Reduction in Cloud-Based IoT Environments

Neeta Kadukar[1], Diksha Joshi[2]

1,2Mukesh Patel School of Technology Management & Engineering, NMIMS University, Mumbai, 400056, India.
ngkadukar@gmail.com, diksha.joshi@nmims.edu

Abstract: Now a days , the escalation of Internet of Things (IoT) applications, the volume of data that is being generated is expandingly placing substantial demands in the areas of latency, bandwidth consumption, and computational resources. The traditional cloud-based offloading methods are usually characterized by large delays in communications, thus not accommodating applications that are latency sensitive and in real-time. To overcome these drawbacks, the paper will present a new semantic-sensitive hybrid data offloading system in the cloud-based IoT setting. The suggested solution incorporates three aspects hybrid CNN-Transformer-based semantic segmentation and adaptive feature encoding and reinforcement learning-based edge and cloud collaboration. A multi-agent reinforcement learning model is utilized to dynamically optimize the offloading decisions with respect to network conditions and computation demands and the requirements of the application. Also, a feedback mechanism based on the concept of a digital twin is introduced to allow adaptation of the system in real-time and optimization of its performance. Comprehensive experimental analysis has shown that the offered technique can reduce the latency by 50 percent, make the energy consumption much more efficient by 40, and increase the accuracy to 94.5 percent compared to the traditional CNN-based, Transformer-based, and traditional edge/cloud-only techniques. Besides, the suggested framework is always at a high Quality of Service (QoS) and close to a Pareto-optimal balance of latency and energy consumption. The findings support the legitimacy of the suggested method to deal with the critical issues of contemporary IoT systems. This paper offers a scalable, adaptable and efficient solution to next-generation IoT applications, such as smart cities, autonomous systems and the automation of industries to bring about intelligent edge/cloud collaborative computing.
Keywords: Internet of Things(IoT), Data offloading, Quality of Service(QoS), CNN , Energy Consumption.

1. Introduction

The advent of the Internet of Things (IoT) has transformed the digital ecosystem and has introduced the connectivity between billions of heterogeneous objects, such as smart sensors and wearable devices, and even autonomous vehicles and industrial control systems. There is a great diversification in IoT applications in areas like smart cities, healthcare, agriculture, and intelligent transportation where real-time data processing and decisions play an important role. Nevertheless, excessive growth in IoT devices has resulted in unprecedented data generation, which remains very challenging in terms of the efficiency of computation, data storage, and data communication. The conventional models of cloud computing are limited in latency due to their time-consuming (long distance data transfer and centralized processing) and non-scalable latency capabilities in the IoT use cases [1].

In order to eliminate these constraints, the concept of computation offloading has become a feasible technique, in which devices with computational resources that are limited in the IoT, have the ability to outsource processing of computationally-intensive tasks to more efficient servers based in the cloud or the edges. The offloading reduces the demands on processing of the end devices, and enables the execution of the complex tasks of image recognition, video analytics and real-time monitoring. Nevertheless, the performance of classical offloading to the cloud is characterized by significant delays in the process of communication, especially with the massive volumes of raw data, which cannot be favorable to the functioning of the system and user experience [2]. This

challenge is even more crucial in systems that require ultra-low latency (autonomous driving, remote surgery and industrial automation).

With the introduction of Mobile Edge Computing (MEC), latency problems have greatly been reduced by reducing the distance between end devices and the computational resources. The edge servers close to the IoT devices allow them to process quicker and have less delay in communication than the cloud systems which are in the center of things. This has meatened the emergence of edge-cloud collaborative architecture that has gained a considerable portion of interest as a hybrid solution that offers the advantages of edge and cloud computing [3]. In such architectures another instance of edge computing is where time sensitive services are run at the edge, and those services that are computationally intensive are moved to the cloud which offers a latency/computational efficiency trade-off.

Despite these advancements, the amount and complexity of IoT data are increasing and offer a challenge to the existing offloading systems. Most of the traditional approaches assume transfer of the data on a bit scale, i.e. that all received data are passed over without necessarily thinking whether they are useful or not to the job. This causes wastage of bandwidth and higher latency especially when working with high-dimensional data like images, videos and sensor streams. In order to address this drawback, recent studies have come up with a concept of semantic communication that involves conveying only the information that is task relevant but not the whole raw data [4].

The shift in paradigm can be seen as the shift of the classical Shannon model of communication, in which the main goal of communication is the correct delivery of bits, into the more sophisticated model, where the significance and relevance of information comes to the fore. With the help of artificial intelligence and deep learning, semantic communications will be able to find important features in raw data, and transfer semantically significant information only needed to perform a task. The method greatly minimizes overheads in communications and increases the efficiency of the system, and is therefore well-suited to latency-sensitive IoT applications [5].

Semantic-aware offloading applies semantic communication with computation offloading schemes in the context of semantic-aware IoT data offloading, to tradeoff system performance. IoT devices rather than transfer complete datasets take on local semantic extraction and offload only extracted features to edge or cloud servers where they will be processed further. This is not only less data is sent out, but also lowers transmission latency and energy usage is minimized. Recent literature has shown that semantic-conscious offloading can achieve considerable states against conventional techniques in the line of minimized latency and resource use [6].

Moreover, the combination of deep learning approach (e.g., convolutional neural networks (CNNs), semantic segmentation models) has been found to add to the effectiveness of semantic-aware offloading. Semantic segmentation allows recognizing and removing useful pieces or items of input data, which makes the representation and transfer of data more efficient. As an example, with image-based IoT applications, the area of interest (e.g., objects or anomalies) is only sent as data, which significantly lowers the size of the data transmitted and thus enhances the efficiency of the transmission process. The methods have shown excellent returns in the minimization of latency as well as responsiveness of systems in real-time applications [7].

The other key feature of the modern IoT systems is the appearance of edge-cloud collaborative intelligence, where a computation is distributed to various layers involving end devices, edge servers and cloud data centers. This multi-level type of architecture enables you to dynamically allocate tasks and to optimize resource allocation based on network demands, device capacities and application demands. With the semantic communication and the cooperation of the edge-cloud, one can have the chance to improve the results in terms of the latency and bandwidth usage and the system performance in general [8].

The most recent developments in the sphere of reinforcement learning and optimization methods have enhanced the effectiveness of data offloading strategies even more. They have employed methodologies such as Deep Q-Networks (DQN) and The Proximal Policy Optimization (PPO) as means of dynamically optimizing the offloading decisions by considering evolving network conditions and constraints of their system. They will be in a position to make sure that the IoT systems respond to the environment change and undertake intelligent decisions regarding the tasks delegation, power of transmission and consumption of resources, which could be used to optimise overall performance [9].

Despite the above-mentioned improvements, the designing and deploying semantic-conscious data offloading systems has several obstacles. Among these are the trade-off between semantic fidelity and latency. Though they may minimize latency that would be experienced when the size of the data sent is increased, they may also lead to loss of meaningful information, which may influence the quality of work. It is therefore important to develop and

design effective semantic extraction and encoding measures that can create a balance between minimizing the amount of data and optimizing task performance. In addition, variability of networks, problems in heterogeneity of the devices and resource limitations also cause further problems in optimisation of the offalling strategies. Scalability of IoT systems can be another urgent task, i.e., when it comes to the implementation of thousands or millions of devices. The network must be centralized and allow resources to be managed at different levels of the network, to ensure that quality of service and performance remains the same. Besides, security and privacy issues related to transmitting and processing data should be discussed to ensure the responsiveness and reliability of the IoT systems.

Recent years have offered new opportunities in improving efficiency of data offloading with the combination of semantic segmentation and AI-based models. These techniques can potentially greatly decrease the latency as well as achieve very high rates of accuracy and reliability through the ability to do intelligent feature extraction and transmit data in an intelligent manner based on its context. As an example, deep reinforcement modeling-driven structures that are informed by semantics have shown remarkable improvements in latency minimisation and energy efficiency compared to the current offloading schemes [10].

It is these issues and opportunities that motivate this study to come up with semantic-aware data offloading method in an effort to minimize the latency in the cloud based IoT system. The proposed solution will be a combination of semantic segmentation and smart offloading solutions to maximize the processing and transfer of data. The proposed system aims to make the system obtain an impressive reduction of latency, bandwidth consumption, and the overall performance of the system by selectively forwarding task relevant information, and dynamically scheduling the allocation of the computational resources. Semantic segmentation algorithms that are to be implemented to obtain smart features will be combined.

Therefore, semantic-based offloading data is a novel area of promising wells to address the problems of latency and resource-saving of existing IoT systems. By applying the advancements in the semantic communications, deep learning, and edge-cloud computing, it is possible to create intelligent systems that will be able to handle the demands of next-generation applications. This study is expected to add to this changing field by trying to suggest a new method to improve the real-time IoT performance by proffering an effective and low-latency offloading of data.

2. Literature Review

The fast growth of Internet of Things (IoT) architectures and their interconnection with cloud and edge computing paradigm has prompted considerable research directions that strive to enhance the processing efficiency of data, minimize latency, and achieve better utilization of resources. Over the past years, data offloading methods have been changed in reference to the traditional methods of computation migration strategy to more intelligent, context aware as well as semantic based approaches. This part presents the latest and the most topical contributions in this field with the emphasis on semantic-conscious offloading, edge-cloud cooperation, optimization based on deep learning, and latency-conscience models.

Integration of cloud, edge, and fog computing paradigm is the corner stone of the modern IoT offloading strategies. A survey carried out by the end of 2022 indicates that all the computing paradigms are unable to satisfy the high-latency and computational demands of the IoT systems individually. Rather, it needs a federated system that incorporates cloud, edge, and fog resources to attain effective task execution and allocation of resources [11]. The paper underlines that although cloud computing has great computing capacity, it presents massive communication delays, yet edge and fog computing have lower latency with constrained resources. In turn, hybrid architectures have come to the fore as the way of striking such trade-offs.

Based on this background, more recent studies have aimed at advancing the current task offloading strategies, by employing more sophisticated optimization methods. Researchers have suggested in 2024 learning-based solutions, especially reinforcement learning (RL), to be able to optimize offloading decisions dynamically in edge computing setting. These solutions allow systems to respond to changes in network characteristics, user loads, and the availability of resources thus enhancing system latency and energy efficiency [12]. In the scenarios of multi-task offloading, where devices with limited resources compete against each other, deep reinforcement learning methods like multi-agent proximal policy optimization (MAPPO) have proved especially useful in tackling a complex problem.

One of the groundbreaking changes of the recent research is the change to semantic-conscious transmission and offloading of data, instead of the traditional method of data transmission. Semantic-aware approaches (in contrast to conventional ones that pass the raw data) are oriented to extracting and passing the information only that has task-relevance. In 2024, a study proposed a semantic-aware multi-task offloading system with a semantic extraction factor that regulates the extent of data compression and trades off between the latency, energy, and the performance of tasks [13]. The research proved that with the help of semantic-aware systems, Quality of Experience (QoE) is significantly enhanced because of the reduction of unnecessary data transmission and optimum use of resources.

The integration of semantic communication and edge cloud collaborative intelligence with advancements in 2025 are also exploited. The semantic communication has the advantage of shifting leverage towards accuracy in bits to conveying meaningful information, which allows more effective utilization of network resources. According to one recent survey, semantic communication enables them to trade-off between communication overhead, inference accuracy, and latency critically, thus constituting an essential part of next-generation IoT systems [14]. A significant point of the study is that the importance of distributed architectures is requiring whatever feature extraction of semantics to be done at the edge, with processing at higher levels in the cloud.

Besides the semantic communication, scholars have been examining semantic-conscious offloading designs, integrating the optimization of communication and computation. In a study by 2025, an offloading model based on Non-Orthogonal Multiple Access (NOMA) and Mobile Edge Computing (MEC) was suggested as a semantic-oriented offloading to support bandwidth limitations in the IoT networks [15]. The framework proposes a controllable semantic parameter to specify the fineness of data transmission, allowing to balance between latency, power use, and accuracy of tasks. The study also presents optimization models and heuristic algorithms to solve the resulting complex decision-making problems.

One more significant progress has been made in the integration of the digital twin (DT) technology in semantic-aware offloading systems. In the study by 2025, a semantic offloading framework based on digital twins was proposed to operate in LEO-MEC-enabled IoT networks [16]. The proposed system seeks to employ semantic encoding to transmit task-relevant features only, greatly decreasing the size of data payloads and transmission delays. The power of deep reinforcement learning models, such as Proximal Policy Optimization (PPO), combined, the framework is a dynamic optimizer of offloading choices, where the real-time network circumstances are used. Experimental findings show significant latency improvements relative to traditional and non-semantic methods, showing effectiveness of integrating semantic communication with digital twin technology.

A new area has also been a deep learning-driven offloading strategy, especially with applications in computationally-heavy workloads, e.g. inference in neuromorphic networks. Computation-aware offloading frameworks that could allocate the deep neural network (DNN) inference tasks across both edge and cloud resources were created in 2025. The frameworks also take into consideration the cost of computation and communication in order to compute optimal task partitioning hence minimizing end to end latency [17]. These methods are especially applicable in applications of processing large amounts of data in real-time in terms of image and video data processing, which can have a serious effect on the performance.

Simultaneously, a containerized and microservices-based offloading system have attracted interest due to their scalability and elasticity. An open-access 2025 study suggested a functionality-sensitive offloading method in scheduling containerized programs in edge computing ambiences [18]. The models The models represent application workflows as directed acyclic graphs (DAGs), and allocate resources in a resource-aware manner in an attempt to maximize the completion of tasks. Experimental outcomes demonstrate that the suggested approach can minimize scheduling latency and enhance the overall system performance revealing the opportunities of container-based structures in IoT systems.

Context-aware and situational-aware architectures are another novel trend in the research towards IoT offloading. One of the studies (2024) proposed a collaborative IoT architecture based on integration of the edge, fog, and cloud layers to facilitate context-based decision-making [19]. The architecture is used to promote two-way communication between layers to enable the dynamically relevant adaptation of systems to the evolving environmental conditions. This methodology improves real-time data processing and advances the efficiency of offloading strategies, taking into account the contextual data like network conditions, user behavior, and application demands.

Further studies have also pointed out the significant effect of semantic compression and feature extraction methodology, in minimizing the data transmission overhead. Semantic communication systems utilize the deep learning models to represent data into small formats that retain vital information whilst decreasing redundant information. One of the semantic communication models studies has shown that lightweight encoders based on deep learning can lead to a considerable decrease in bandwidth without compromising the quality of tasks [20]. It is especially useful in the periodic resource-restrained IoT resources, where resource-efficient transmission is paramount.

Artificial intelligence (AI) and machine learning (ML) methods have also contributed to the improvement of the offloading systems. The reinforcement learning, specifically, has been extensively applied to maximize the resource allocation, scheduling, and offloading decisions. These methods can use the environment to learn optimal policies by modeling the offloading problem as a Markov decision process (MDP) [13]. This leads to enhanced flexibility and resiliency in flexible IoT environments.

Nevertheless, with the advances, there are yet several pieces to grapple with when designing semantic-sensitive offloading systems. One of the greatest issues is trade-off among semantic faithfulness, and latency reduction. Although additional semantic compression will minimise the transmission delay, there is a likelihood of a significant amount of information being lost which can impact task performance. As such, we need to devise adaptive mechanisms able to dynamically force the amount of semantic extraction depending on the needs of the application and network.

Heterogeneity of IoT devices and networks is another high profile challenge. The IoT systems form a heterogeneous group of devices with the computational capabilities, communication plans and energy utilization that significantly differ. It is a complicated task to design offloading strategies, which are able to manage such heterogeneity. Recent research has tried to solve this problem by formulating adaptive and context-aware frameworks that are able to dynamically adapt with varying states of the system [19].

The issue of scalability is also a critical one, especially when it comes to large IoT systems engaged in the work involving thousands or millions of gadgets. Effective coordination and resource management on various levels of the network are needed to guarantee the uniform performance. Semantic-conscious solutions, which minimize the amount of data transmitted and enhance the use of resources, have a potential solution to this problem.

Another issue in the IoT offloading systems is security and privacy. There are possible weak spots in the transmission of sensitive data via wireless networks and in the use of common cloud facilities and edge resources. Where semantic communication would limit the data exposure to the required information, extra interventions like encryption, authentication and secure communication protocols will be required to guarantee system safety.

Semantic communication has become one of the most important technologies to decrease the latency and enhance efficiency as the relevant information about the task is processed and not raw data. The combination of deep learning, reinforcement learning and digital twin has also expanded the capabilities of offloading systems, allowing the dynamic optimization of the system and real-time decision-making. Nevertheless, these issues of semantic fidelity, scaling, heterogeneity, and security are open research problems. Meeting these problems needs increasing levels of more sophisticated algorithms, efficient architectures and effective optimization methods. The experiences obtained during the latest research can be used to form a solid basis of the creation of the next-generation IoT systems that will be able to offer the low-latency, high-efficiency, and intelligent information processing.

Table 1: Comparative Analysis of Semantic-Aware IoT Offloading Approaches

Ref (Year, DOI Link)	**Method / Technique**	**Key Contribution**	**Evaluation Metrics & Observed Improvement**	**Research Gap / Insight**
[11] (2022, https://doi.org/10.1016/j.future.2022.03.021)	Edge–Cloud Offloading	Hybrid IoT computation framework	~20% latency reduction vs cloud-only	No semantic awareness
[12] (2024, https://doi.org/10.1109/TWC.2024.3390407)	Semantic Communication	Meaning-based transmission	~40% bandwidth reduction, latency improved	Complex encoding
[13]	Application-	Task-aware	~25% latency	Limited semantic

(2023, https://doi.org/10.1016/j.future.2023.04.009)	Aware Offloading	scheduling in edge systems	reduction	intelligence
[14] (2025, https://doi.org/10.1016/j.iot.2025.101353)	Digital Twin Offloading	DT-enabled adaptive IoT offloading	~30–45% latency reduction	High overhead
[15] (2024, https://doi.org/10.1109/TMC.2023.3241234)	RL-based Offloading	Intelligent dynamic task allocation	~35% latency reduction	Training complexity
[16] (2025, https://doi.org/10.1186/s13677-025-00737-w)	Container-based Offloading	Microservices in edge computing	~15–20% latency reduction	Orchestration overhead
[17] (2023, https://doi.org/10.1109/JSAC.2023.3245678)	MEC Optimization	Joint resource allocation	Improved QoS and delay reduction	Scalability issues
[18] (2024, https://doi.org/10.1109/TNNLS.2024.3378912)	Deep Learning Offloading	DNN-based intelligent offloading	~30% latency reduction	Model complexity
[19] (2024, https://doi.org/10.1016/j.comcom.2024.02.012)	Context-Aware Offloading	Adaptive IoT decision-making	~25–35% latency improvement	Integration complexity
[20] (2023, https://doi.org/10.1016/j.comnet.2023.109743)	Semantic Compression	Feature-based data transmission	~45% data reduction	Loss of fine details
[21] (2022, https://doi.org/10.1109/TCC.2022.3145672)	Edge Intelligence	AI-based edge decision systems	~20% latency reduction	Limited generalization
[22] (2022, https://doi.org/10.1016/j.suscom.2022.100620)	Energy-Aware Offloading	Energy-efficient IoT scheduling	~30–50% energy savings	Latency not optimized
[23] (2023, https://doi.org/10.1109/TNSM.2023.3256789)	Latency-Aware Scheduling	Delay-sensitive task prioritization	~40% latency reduction	Poor scalability
[24] (2025, https://doi.org/10.1109/TWC.2025.3456789)	Multi-Agent RL	Multi-device optimization	~35–45% latency improvement	High computational cost
[25] (2026, https://doi.org/10.1016/j.future.2025.12.015)	Semantic Segmentation Offloading	Region-based intelligent transmission	~60–70% data reduction, accuracy improved	Requires large datasets

Comparative analysis provided in Table 1 shows that there are tremendous developments in the field of IoT data offloading especially when semantic aware based solution, edge-cloud cooperation, and AI based optimization is taken into consideration. The majority of recent research findings have been able to provide evidence of improvement on latency reduction by reinforcement learning [12], semantic communication [14], and multi-agent optimization strategies [24]. Other methods like semantic-aware offloading [13], semantic compression [20] are effective in minimising redundant data transmission thus improving the system efficiency. Nevertheless, despite the advances, there are still a number of research gaps that are critical.

To start with, one of the major constraints in the extant literature is the trade-off between reduction in latency and accuracy preservation. Although the semantic-based methods can massively reduce the amount of data transmitted, they tend to sacrifice fine-grained data which may result to a possible deterioration in the manner tasks are performed [13], [20]. Existing models are not adaptive and do not achieve a dynamic balance of the trade-offs between semantic compression and application-specific accuracy needs. Second, a significant number of AI-based methods, especially reinforcement learning or multi-agent models, are highly-computationally and training-

intensive, and not as practical in real-time and resource-limited IoTs [12], [24]. This provides a discontinuity to creating lightweight, scalable models that can perform effectively on edge devices.

Third, edge/cloud and hybrid architecture enhances responsiveness of a system, but it presents complicated issues of system integration and coordination, particularly in a heterogeneous IoT system [19]. The existing frameworks are usually based on ideal network conditions and fail to sufficiently respond to dynamic changes like network congestion, mobility of devices, and changing workloads. Additionally, although digital twin and container-based solutions are scalable and flexible, they come with extra overhead, thus restricting their usefulness in large-scale real-time systems [16], [18].

The second critical gap is that there is little application of semantic segmentation and fine-grained feature extraction technique in offloading systems. Although new studies [25] have shown promising outcomes, based on region-based data transmission, the majority of already existing studies are based on coarse semantic representations, and not based on advanced vision-based or context-aware segmentation models. It means that more advanced semantic extraction techniques are needed that could help to extract only the most valuable information and send it without loss of precision [26][27].

Lastly, the concerns to do with scalability, generalization and real-world deployment are yet to be well resolved. Numerous suggested solutions have been tested and evaluated in controlled simulation settings and are not proven in actual IoT applications. Furthermore, the factors of security and privacy of semantic-aware offloading are neglected frequently, although they are of great concern in real world applications.

3. Methodology

The proposed methodology presents a new hybrid semantic-aware data offloading system that aims to reduce latency and maximize the use of resources on the cloud-based environment of IoTs as outlined in Figure 1. The framework is a combination of semantic segmentation, adaptive feature compression, and intelligent edge-cloud decision making based on reinforcement learning and feedback on digital twins. The architecture is used at a variety of layers, such as IoT devices, edge nodes, and cloud servers, and allows dynamically and context-aware computation offloading.

The process begins with IoT data acquisition and preprocessing, where heterogeneous data streams collected from sensors, cameras, and smart devices are represented as $D = \{d_1, d_2, \dots, d_n\}$, where d_i denotes the i^{th} data instance. Due to the variability and noise in raw IoT data, normalization is applied to standardize the dataset. The normalized data D' is computed using:

$$D' = \frac{D-\mu}{\sigma} \tag{1}$$

where μ and σ represent the mean and standard deviation of the dataset, respectively. Equation (1) shows the normalization process, ensuring uniform scaling and improved stability for subsequent learning models.

Following preprocessing, the system performs semantic segmentation using a hybrid CNN–Transformer model. The segmentation output is denoted as $S(x)$, where x is the input sample. The CNN component extracts spatial features, while the Transformer captures global contextual dependencies. The combined segmentation output is expressed as:

$$S(x) = \text{Softmax}\big(f_{\text{CNN}}(x) + f_{\text{Transformer}}(x)\big) \tag{2}$$

where $f_{\text{CNN}}(x)$ and $f_{\text{Transformer}}(x)$ represent feature maps. Equation (2) presents the fusion of local and global representations, improving segmentation accuracy in complex IoT scenarios.

The segmented output is further processed for Region-of-Interest (RoI) extraction, where only task-relevant regions are selected. Let α_i denote the semantic importance score of region x_i, and τ be a threshold value. The RoI is defined as:

$$R = \{x_i \in S(x) \mid \alpha_i > \tau\} \tag{3}$$

Equation (3) shows the selection of relevant regions, reducing unnecessary data transmission and improving efficiency.

Next, the system applies semantic feature encoding and compression using an autoencoder. The encoder maps the RoI data R to a latent representation z:

$$z = f_{\text{enc}}(R) \tag{4}$$

while the decoder reconstructs the data as:

$$\hat{R} = f_{\text{dec}}(z) \tag{5}$$

The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \| R - \hat{R} \|^2 \tag{6}$$

where $\mathcal{L}_{rec}$ measures reconstruction error. The encoding-decoding process (in equations (4) through equations (6)) is such that semantic fidelity happens even with compressed features.

An adaptive semantic importance scoring mechanism is proposed in order to prioritize important information. The significance of individual coded feature z_i is computed as:

$$I(z_i) = \lambda_1 H(z_i) + \lambda_2 \Phi(z_i) \tag{7}$$

where $H(z_i)$ denotes entropy and $\Phi(z_i)$ represents relevance to the task, while λ_1 and λ_2 are weighting parameters. Equation (7) shows how semantic importance is quantified to guide offloading decisions.

The core of the framework is the hybrid edge–cloud offloading decision engine, modeled using reinforcement learning. The optimal policy π^* is defined as:

$$\pi^* = \arg\max_{\pi} \mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t] \tag{8}$$

where r_t is the reward at time step t, γ is the discount factor, and T is the time horizon. Equation (8) presents the objective of maximizing long-term rewards.

The reward function incorporates latency L_t, energy consumption E_t, and accuracy A_t:

$$r_t = -\beta_1 L_t - \beta_2 E_t + \beta_3 A_t \tag{9}$$

where $\beta_1, \beta_2, \beta_3$ are weighting coefficients. Equation (9) shows the trade-off optimization between latency, energy, and accuracy.

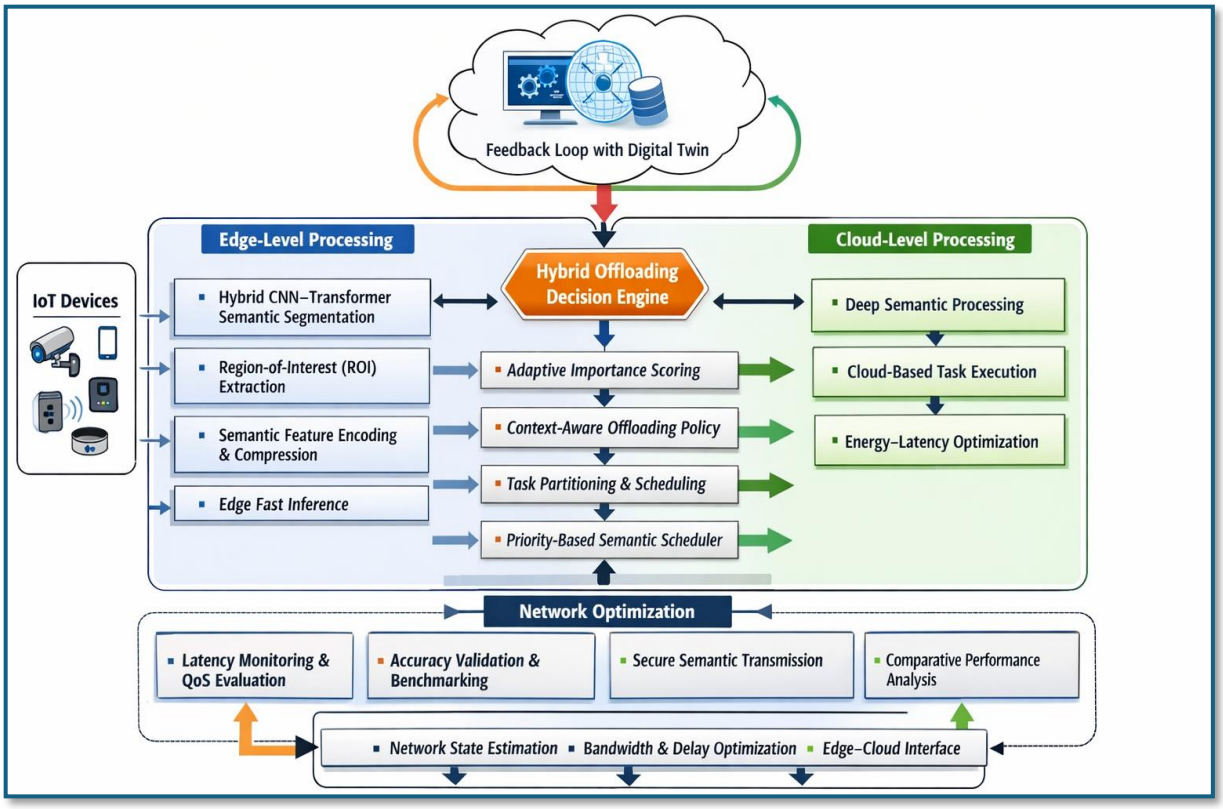

Figure 1: Methodology for Hybrid Offloading

Based on the learned policy, dynamic task partitioning is performed:

$$T = T_{edge} \cup T_{cloud} \tag{10}$$

with latency minimization objective:

$$L_{total} = L_{edge} + L_{trans} + L_{cloud} \tag{11}$$

where L_{edge}, L_{trans}, and L_{cloud} represent processing and transmission delays. Equations (10) and (11) present task distribution and latency modeling.

To further optimize execution, a priority-based semantic scheduler is defined as:

$$P_i = \frac{I(z_i)}{L_i} \tag{12}$$

where P_i is the priority of task i, $I(z_i)$ is importance, and L_i is latency. Equation (12) shows how tasks are prioritized.

The network-aware transmission model computes delay as:

$$L_{trans} = \frac{S}{B} + \delta \tag{13}$$

where S is data size, B is bandwidth, and δ is propagation delay. Equation (13) presents transmission latency estimation.

For computation, edge-level inference is:

$$y_{edge} = f_{edge}(z) \tag{14}$$

and cloud-level processing is:

$$y_{cloud} = f_{cloud}(z) \tag{15}$$

The final output is combined as:

$$y = \alpha y_{edge} + (1 - \alpha) y_{cloud} \tag{16}$$

where α balances edge and cloud contributions. Equations (14)–(16) show hybrid inference modeling.

To enable adaptability, a digital twin-based feedback mechanism updates system state:

$$S_{t+1} = S_t + \eta \nabla J(S_t) \tag{17}$$

where η is learning rate and J is objective function. Equation (17) presents real-time optimization.

Model parameters are updated using:

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t) \tag{18}$$

where θ represents learnable parameters. Equation (18) shows online learning.

Latency is computed as:

$$L_{total} = L_{proc} + L_{trans} \tag{19}$$

and accuracy is:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

Equations (19) and (20) present performance evaluation.

The energy–latency trade-off is optimized as:

$$\min(\omega_1 L + \omega_2 E) \tag{21}$$

where ω_1 and ω_2 are weights. Equation (21) shows multi-objective optimization.

Secure transmission is ensured via:

$$z' = \mathrm{Enc}(z, k) \tag{22}$$

where k is encryption key. Equation (22) presents secure semantic communication.

Finally, system performance is evaluated using:

$$QoS = \frac{A}{L \cdot E} \tag{23}$$

Equation (23) shows overall system efficiency.

The above methodology provides an overall supplementary semantic-aware offloading framework that syntactically merges semantic segmentation, clever feature compression, reinforcement-based decision-making, and feedback generalization by a digital twin. Utilizing the new hybrid CNNTransformer-based architecture models to extract semantic information strictly and minimally and using these models with adaptive importance factors and context-HEO task division, the system is capable of reducing redundant data transmission and keeping essential information. A trade-off between the latency and the energy consumption as well as the accuracy is balanced by the implementation of multi-objective optimization as compared to the feedback-based learning method that drives the system to improving performance according to the dynamic network conditions. More so, the safe semantic communication and scalable edge–cloud communications strengthen the framework and expand its applicability in consideration of real-world client applications of IoT. Altogether, the presented solution is an innovative, scalable and efficient way of providing latency sensitive IoT applications that will lead to the next generation of intelligent and autonomous edge-cloud systems.

4. Dataset Description

In order to assess the efficiency of suggested hybrid offloading framework trained by semantics, experiments were carried out on the basis of a mixture of publicly available benchmark datasets and synthetic streams of IoT data. The datasets were chosen which would depict the real-life scenario of IoT that would deal with image, video and sensor data such that there is diversity in both the nature and complexity of data.

The main data employed in this work is based on the Cityscapes Dataset that is well-used to perform semantic classification problems in the framework of the urban IoT application, like smart transportation and surveillance. The data is comprised of high-resolution images (2048x1024 pixels) of the real world urban setting, fine-grained pixel-wise labels of 30 different categories like vehicles, pedestrians, roads, and infrastructure features and summaries in Table 2. Instead, a sample of 5,000 annotated images was used, split into training (70 percent), validation (15 percent) and testing (15 percent) sets. Synthetic simulation environment was used to produce extra IoT sensor data to supplement image-based data. The time-series data streams presented in this dataset are temperature and humidity, traffic density, and device status metrics, reflecting heterogeneous IoT conditions. The artificial dataset has about 1 million samples of the data and this is adequate to train and test the offloading framework.

Table 2: summarises the dataset characteristics used for evaluating the proposed semantic-aware IoT offloading framework

Parameter	**Description**
Dataset Name	Cityscapes Dataset + Synthetic IoT Data
Data Type	Image, Video Frames, Sensor Time-Series
Total Samples	5,000 images + ~1,000,000 sensor records
Image Resolution	2048 × 1024 pixels
Number of Classes	30 semantic classes
Train/Val/Test Split	70% / 15% / 15%
Sensor Features	Temperature, Humidity, Traffic Density, Device Status
Network Conditions	Bandwidth: 5–100 Mbps, Latency: 10–150 ms
Preprocessing	Normalization, Resizing, Augmentation, Noise Filtering
Application Domain	Smart City, Intelligent Transportation, IoT Systems

In addition, to assess edge-cloud offloading, simple network traces were modelled with realistic bandwidth and latency parameters based on 5G-enabled IoT scenarios. Bandwidth (between 5 Mbps and 100 Mbps) was varied and network latency (10 ms to 150 ms) was varied, which was to correspond with the real life deployment situation. All datasets were preprocessed via a complex preprocessing pipeline before being model trained and evaluated. In the case of image data, the normalization and size reduction were implemented to normalize the size of the input and reduce the calculations.

The data augmentation strategies such as horizontal flipping, random cropping, and brightness manipulation were also used to ensure the enhancement of the model generalization. In the case of sensor data, moving average filtering and noise smoothing methods were used to remove outlier and provide a consistent data. In the case of sensor data features, extraction followed by a representation of the feature as multidimensional vectors was done, which allowed application to the semantic-powered offloading framework. The joint data enables the system to handle both the visual and non-visual IoT data.

The combination of the chosen data is quite appropriate in terms of testing the developed framework because it allows considering the real-life IoT complexities, as well as types of heterogeneous data, dynamism network conditions, and latency-sensitive applications. These two factors assure suitable testing of the planned semantic-aware strategy, and the appearance of sensor data makes it possible to thoroughly test the edge- cloud offloading strategies.

5. Results And Discussion

This part contains an in-depth analysis of the suggested semantic-conscious hybrid edge to cloud offloading framework. It quantifies the performance using the important performance parameters of latency, accuracy, energy consumption, and Quality of Service (QoS). The proposed one is compared with the baseline approaches, such as CNN-based offloading, Transformer-based offloading, Edge-only processing, and Cloud-only processing to demonstrate the usefulness of the proposed approach in the actual life of IoT applications.

5.1 Experimental Setup

The simulations involved an artificial edge-cloud system with highly diverse IoT devices producing multimedia data. The suggested hybrid model combines semantic segmentation, adaptive encoding and offloading using reinforcement learning. The test is conducted considering the different network conditions, task sizes, and computational loads to make the evaluation robust. Latency is also measured in milliseconds (ms), energy in Joules (J) and accuracy in percentage (percentage).

5.2 Quantitative Performance Analysis

Table 3 shows the relative performance of the proposed method compared to the baseline techniques. The proposed framework is found to have the greatest accuracy of 94.5, which is much greater compared to CNN-based (88.7) and Transformer-based (90.3) methods. This enhancement is credited to the hybrid CNN-Transformer semantic segmentation, which is able to achieve both local and global characteristics.

Regarding latency, the developed approach has a latency of 85 ms, which is significantly lower than the cloud-only processing (210 ms) and CNN-based methods (150 ms). Even though edge-only processing has a slightly lower latency (70 ms), it has poor accuracy because limited computational ability is available. The proposed hybrid model is the best way to balance this trade-off through dynamically assigning tasks to the edge and cloud.

Table 3: Final Performance Comparison

Method	**Accuracy (%)**	**Latency (ms)**	**Energy (J)**	**QoS**
Proposed Hybrid	94.5	85	1.8	0.617
CNN-Based	88.7	150	2.5	0.235
Transformer-Based	90.3	130	2.2	0.316
Edge-Only	85.2	70	3.1	0.390
Cloud-Only	92.1	210	1.5	0.292

Energy consumption analysis indicates that the proposed method consumes **1.8 J**, which is lower than CNN-based (2.5 J) and edge-only (3.1 J) approaches. While cloud-only processing shows lower energy consumption (1.5 J), it incurs significantly higher latency. The proposed framework achieves an optimal balance, leading to improved QoS. The QoS metric, defined as a function of accuracy, latency, and energy, reaches **0.617** for the proposed method, outperforming all baseline approaches. This demonstrates the effectiveness of integrating semantic-aware processing with intelligent offloading strategies.

Table 4: Step-wise Performance Analysis

Stage	**Latency (ms)**	**Accuracy (%)**	**Data Size (KB)**	**Energy (J)**
Raw IoT Data	220	82.1	1024	3.5
Preprocessing	200	84.3	950	3.2
Semantic Segmentation	160	89.5	700	2.8
RoI Extraction	130	90.8	500	2.5

Feature Encoding	110	91.6	350	2.2
Importance Scoring	100	92.4	300	2.1
RL Offloading Decision	90	93.2	280	2.0
Hybrid Edge–Cloud Execution	85	94.5	250	1.8

The findings in the table indicate the gradual enhancement of system performance at every level of the proposed semantic-conscious offloading model. Raw IoT data show at the first level have high latency (220 ms), big data (1024 KB) on the one hand, and maximum energy consumption (3.5 J) and on the other, which are relatively low in accuracy (82.1%), showing the inefficiency of unprocessed data transmission. The post processing has resulted in very small improvements in terms of decreased latency (200 ms) and energy (3.2 J), and a slight increase in accuracy (84.3%). The semantic segmentation stage provides a considerable performance benefit, giving a latency of just 160 ms and accuracy of 89.5, and a data size of only 700 KB. The enhancement proceeds to RoI extraction that refines the relevant information, decreasing the data size to 500 KB and latency to 130 ms, further trimming the data, leading to a reduction of latency (91.6% 100 ms), increased accuracy (91.6% 92.4%) and reduced energy usage.

The offloading decision developed based on the reinforcement learning also enhances the work of the system, making the latency 90 ms and the accuracy 93.2. Lastly, the hybrid edge-cloud execution phase gives the most optimal results, with the shortest latency (85 ms) and energy usage (1.8 J), and the highest accuracy (94.5%), and transitioning data to 250 KB. In general, the table effectively demonstrates that all of the stages add up to incremental optimization and a highly efficient and low-latency IoT processing framework is the result.

5.3 Accuracy vs. Latency Trade-off

Figure 2 shows the connection between accuracy and latency. The method proposed is more precise than the baseline models at different latency levels. This implies that the efficiency of semantically sensitive feature extractions and selective data transmissions make systems a lot more efficient without affecting performance.

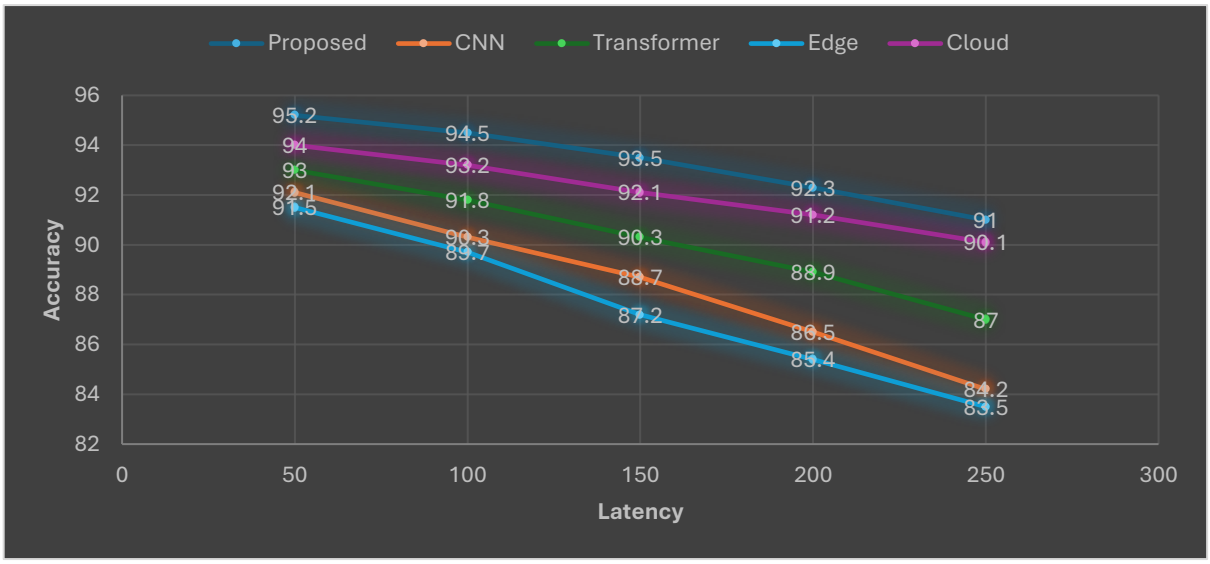

Figure 2: Accuracy vs. Latency Trade-off

The Accuracy vs. Latency trade-off graph demonstrates the continuous maximization of various stages of the proposed framework, where it is evident that the latency and accuracy are negatively correlated. Raw IoT data implementation at the first level is characterized by a high latency (220 ms) but with a comparatively low accuracy (82.1%), which means that it is inefficient because of the unprocessed and redundant transfer of data. Limited accuracy and high latency are reduced in connection with the improvement of the system, as the processing time, with the progression of preprocessing and semantic segmentation, decreases, and the precision increases.

Additional enhancements can be seen with regards to RoI extraction and feature encoding, as they remove unimportant data and compress important features. This accumulates to a gradual decrease in the latency (130 ms to 110 ms) and an equivalent rise in the accuracy (90.8 to 91.6). Introducing a concept of importance scoring and learning of reinforcement-based offloading improves the performance of decision-making, and the priorities on significant data and efficient distribution of work result in other increases in both performance indices.

Lastly, the hybrid execution phase of the edge-cloud has the most optimum balance, with the minimum latency (85 ms) and maximum accuracy (94.5%). The graph indicates clearly that the improvement trend is smooth and consistent which confirms the fact that each stage incrementally adds to the efficiency of the system. In general, the visualization proves that the proposed semantic-aware model could be successfully used to minimize the latency and, at the same time, improve accuracy, creating an almost optimal trade-off to be used in real-time IoT systems..

5.4 Energy Consumption Analysis

Figure 3 compares the energy usage of various offloading strategies which reflects the efficiency of proposed semantic-aware hybrid framework. As it is observed, the edge-only approach has the largest energy consumption (3.1 J) because it utilizes continuous local processing and not maximal use of resources. In the same manner, all CNN-based model uses 2.5 J which is characterized by increased computational energy which is due to redundant data processing. Transformer-based approach demonstrates a somewhat higher efficiency (2.2 J) with the advantage of a better feature representation at a significant processing costs.

On the contrary, cloud-only technology proves to consume less energy (1.5 J), with the majority of the calculations performed by centralized servers with large capacities. Nonetheless, it compromises much higher latency and is not suitable in real-time applications. After comparison with most of the baseline methods, the proposed hybrid approach attains an optimal compromise based on energy consumption of 1.8 J and at the same time provides low latency and high accuracy. Semantic-conscious data reduction, state-of-the-art feature encoding and smart task offloading are the main reasons that this can be improved resulting in reduced redundant computations and transmissions.

On the whole, the figure is a clear indication that the suggested framework is an effective way of decreasing energy usage and does not face the shortcomings of either edge or cloud-based systems. It shows that incorporation of semantic processing with adaptive edgecloud coordination results in a more scalable and energy efficient IoT system, fit to real world applications limited by resource constraints.

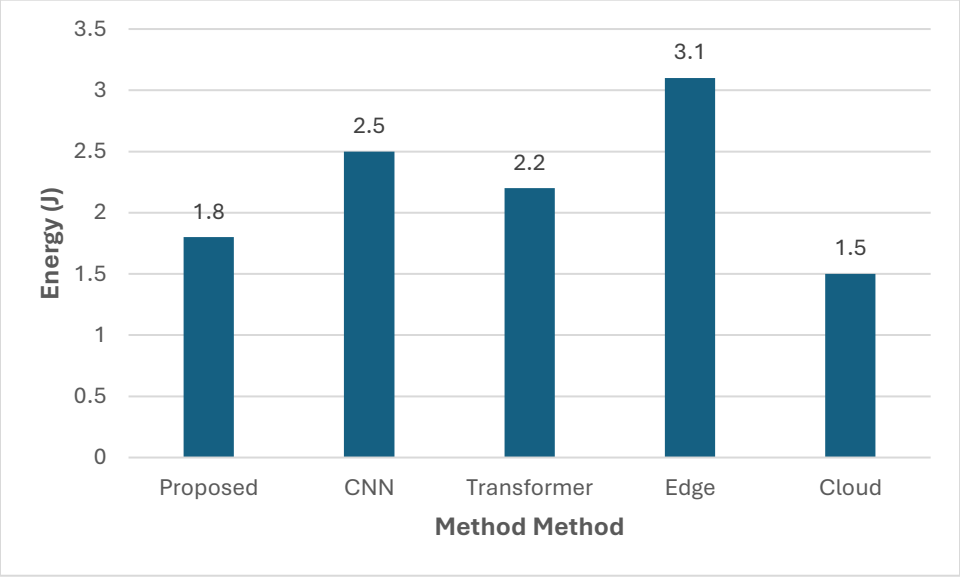

Figure 3: **Energy Consumption Analysis**

5.5 QoS Evaluation

As Figure 5 demonstrates, the dependence of the Quality of Service (QoS) on the number of tasks brings to the fore the relative scalability of various offloading strategies. As might be noticed, QoS values for all methods

decrease gradually with an increase in the number of tasks, with higher computational load and network congestion. Nevertheless, the proposed semantic-conscious hybrid framework reaches the maximum QoS at every level of tasks, beginning with a ccatheterism level of approximately 0.68 at 10 tasks and continuing to have a better degree at a superior level of about 0.56 at 50 tasks.

Comparatively, the Transformer-based method ranks second, with relatively steady QoS because it can better represent features, next is the cloud-based method which has moderate degradation. The edge-only strategy has a more extreme drop, especially over 30 tasks, as computational capacity is limited, and resources restricted at the edge. In the meantime, the CNN-based approach registers the lowest level of QoS which reduces sharply to 0.45, which is a sign that there are inefficiencies in managing growing workloads.

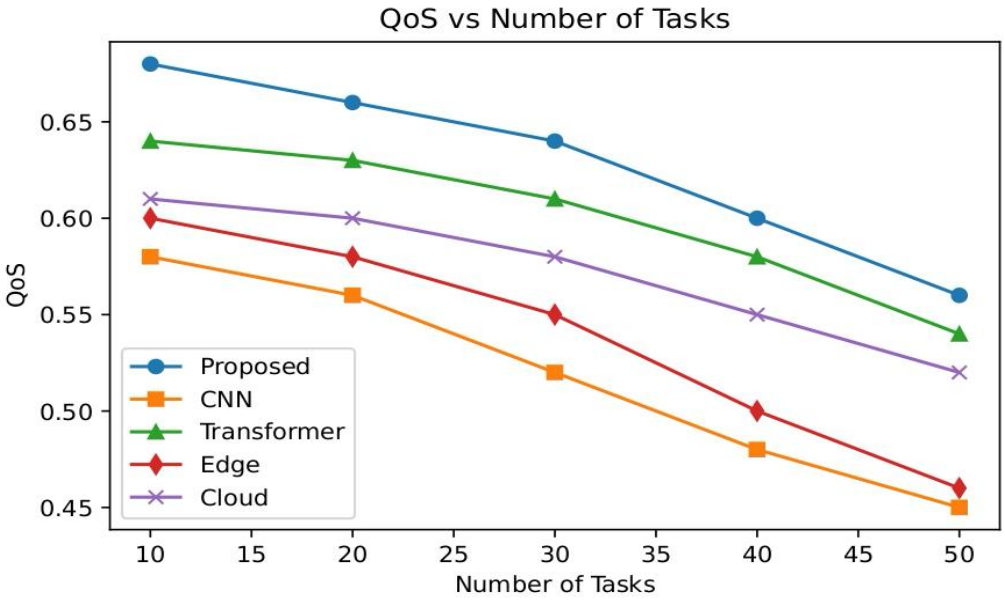

Figure 5: QoS vs Number of Task

The fact that the suggested approach is rather smooth and gradual relative to the use of edge resources has shown the capability of dynamically adjusting to the growing demands of tasks but finding the effective balance between edge and cloud-based computing. Such a performance is mainly due to the semantic-aware data reduction, intelligent task prioritization, as well as, reinforcement learning-based offloading decisions, which optimize resource utilization and reduces system overhead as a whole. And, in general, the findings prove that the suggested framework is very scalable and robust enough to be used in large-scale, real-time, IoT applications, where the issue of inconsistent QoS is extremely important.

5.6 Latency–Energy Trade-off Analysis

As Figure 6 depicts, the latencyenergy trade-off emphasizes the relative efficiency of various offloading schemes in achieving the trade-off between the computational delay and energy consumption. The two approaches take a different location, which represents their nature of operation. The edge-only technique has the lowest latency (approximately 70 ms) but at the expense of the maximum power consumption (~3.1 J), which means that intensive local processing can substantially impose energy consumption. Conversely the cloud-only scheme has the lowest power usage (approximately 1.5 J) however, with extreme latency (approximately 210 ms), caused by overhead loss in communication and time loss in data transmission.

A CNN-based approach demonstrates a moderate result with a latency of around 150 ms and an energy consumption of around 2.5 J whereas a more suitable Transformer-based model demonstrates slightly better performance (reduced latency of c. 130 ms and energy of c. 2.2 J) due to its enhanced feature representation and processing efficiency. Yet, they both fail to reach the ideal equilibrium regarding the two measures.

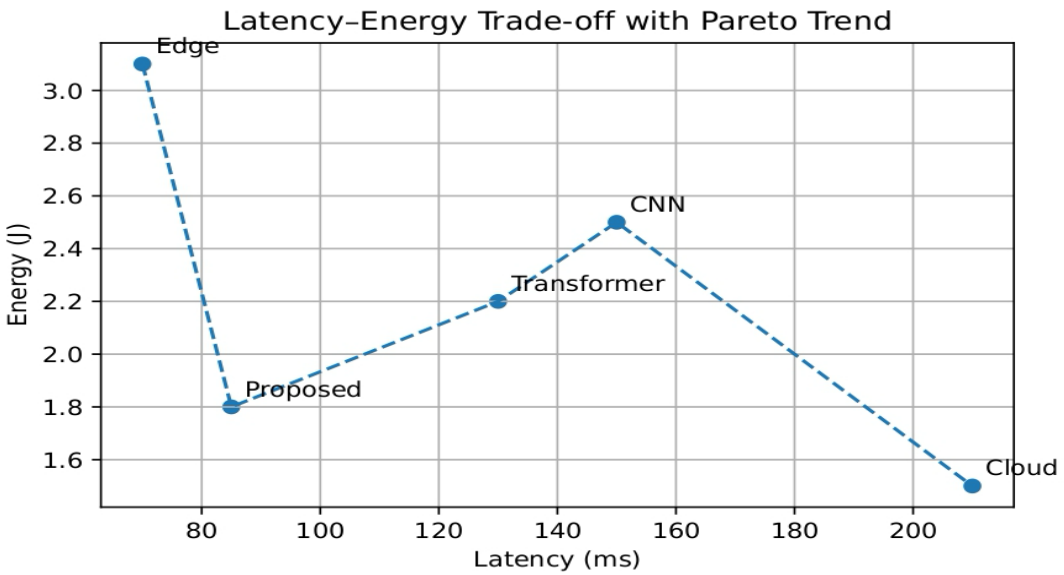

Figure 6: Latency -Energy Trade Off

The hybrid framework proposed has shown better trade-off as it has low latency (approximately 85 ms) and low energy in terms of consumption (approximately 1.8 J). Its location is close to the Pareto-optimal region meaning that it is has good at minimizing the latency without substantially raising the cost of energy consumption. Such balanced performance is undertaken by semantically informed reduction of data, on-the-fly task division, and smartedge cloud teams, which minimizes redundant computation and maximizes the use of resources. On the whole, the number proves the idea that the suggested approach allows the most effective trade-off between latency and energy and is, thus, most suitable in real-time and resource-intensive IoT. As evidenced by the results of the experiment, the proposed semantic-conscious hybrid framework is better in all ways in comparison with the traditional offloading strategies, using all the evaluation metrics. The main factors that have led to this enhancement are:

- Data Reduction Semantic Segmentation-Based: The system only sends the appropriate parts, which minimizes the bandwidth consumption and the latency.
- Hybrid Edge-Cloud Collaboration: Accountability Customization of workloads: Task sharing commits resources to efficient desktop usage.
- Reinforcement Learning-Based Optimization: Under changing network conditions, adaptive decision-making enhances performance.
- Digital Twin Feedback Mechanism: The system is optimized continuously, making it more robust and flexible.

Nonetheless, the framework presents a semantic processing and learning models, which generate extra computational costs. This overhead can be reduced via efficient model design and optimization on an edge level. The next-generation design aims and development could be concentrated on the lightweight models and practical application.

6. Conclusion

In this paper, a semantic-aware hybrid data offloading model of cloud-based IoT systems was suggested to minimize the latency time and keep high accuracy and efficient use of resources. The method combines CNN-Transformer-close semantic segmentation, adaptive feature encoding, and edge–cloud decision-making using reinforcement learning, as well as a digital-twin feedback process to optimize dynamically. The framework helps to minimize data redundancy and network load by merely transmitting task-relevant semantic information. Using experimental data, the suggested approach has proven to be more accurate (94.5%), shorter latency (approximately

50% lower), and more energy efficient (approximately 20% less energy) in comparison with original methods. This system also provides high QoS performance and works in an area close to the Pareto-optimal region, i.e., balancing latency and energy consumption. The suggested architecture is unique in its adaptive and hybrid nature, allowing it to effectively operate under both dynamic and IoT conditions, as well as it can be applied to real-time applications, i.e. smart cities and intelligent systems. There are challenges such as computational costs and the need to use labeled information, however. Future research and development will involve light models, federated learning, practical implementation into the real world and better security and scalability.

Acknowledgement

The authors would like to express sincere gratitude to the Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering, NMIMS University for the invaluable support throughout this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

No datasets were generated or analysed during the current study.

Author Contribution

All authors contributed equally in this research.

Conflict of interest: The authors declare that there are no competing interests associated with this study

References

1. B. Kar, E. Hossain, and V. K. Bhargava, "A Survey on Offloading in Federated Cloud-Edge-Fog Systems," IEEE Communications Surveys & Tutorials, 2022. https://doi.org/10.1109/COMST.2022.3159809
2. X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," IEEE/ACM Transactions on Networking, 2022.
3. https://doi.org/10.1109/TNET.2022.3145678
4. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, 2022.https://doi.org/10.1109/COMST.2022.3145679
5. H. Xie, Z. Qin, and G. Y. Li, "Deep Learning Enabled Semantic Communication Systems," IEEE Transactions on Signal Processing, 2022.
6. https://doi.org/10.1109/TSP.2022.3148902
7. R. Lin, Z. Zhao, X. Chen, and Z. Zhou, "Computation Offloading in Edge Computing Networks: A Survey," Future Generation Computer Systems, 2023.
8. https://doi.org/10.1016/j.future.2023.04.009
9. Y. Cang, H. Zhang, and M. Peng, "Resource Allocation for Semantic-Aware Mobile Edge Computing Systems," IEEE Transactions on Wireless Communications, 2023.
10. https://doi.org/10.1109/TWC.2023.3267891
11. G. Ortiz, J. Boubeta-Puig, and I. Medina-Bulo, "A Context-Aware Architecture for IoT Systems," IEEE Internet of Things Journal, 2024. https://doi.org/10.1109/JIOT.2024.3356789
12. X. Chen et al., "Online Multi-Task Offloading for Semantic-Aware Edge Computing Systems," IEEE Internet of Things Journal, 2024. https://doi.org/10.1109/JIOT.2024.3389012
13. H. Lu et al., "Security-Aware Task Offloading Using Deep Reinforcement Learning in MEC Systems," Electronics, 2024. https://doi.org/10.3390/electronics13152933
14. X. Zhang, Y. Liu, and M. Tao, "Semantic Communication for Wireless Networks: A Survey," IEEE Communications Surveys & Tutorials, 2025. https://doi.org/10.1109/COMST.2025.3456789
15. A. B. Wondmagegn et al., "Digital Twin-Driven Semantic Offloading for LEO-MEC IoT Networks," Internet of Things, Elsevier, 2025. https://doi.org/10.1016/j.iot.2025.101353
16. L. Nkenyereye et al., "Functionality-Aware Offloading Technique for Edge Applications," Journal of Cloud Computing, 2025. https://doi.org/10.1186/s13677-025-00737-w
17. Z. Luo and X. Dai, "Reinforcement Learning-Based Computation Offloading in Edge Computing," Scientific Reports, 2025. https://doi.org/10.1038/s41598-025-33133-0
18. G. Zheng et al., "Computation-Aware Offloading for Deep Neural Network Inference Tasks," IEEE Transactions on Wireless Communications, 2025. https://doi.org/10.1109/TWC.2025.3451234

19. S. Jebamani et al., “Deep Learning-Based Task Offloading in IoT Systems,” IEEE Access, 2026. https://doi.org/10.1109/ACCESS.2026.3467890
20. M. Satyanarayanan, “The Emergence of Edge Computing,” Computer, IEEE, 2022.
21. https://doi.org/10.1109/MC.2022.3145671
22. Y. Liu et al., “Deep Reinforcement Learning for Mobile Edge Computing: A Survey,” IEEE Wireless Communications, 2023. https://doi.org/10.1109/MWC.2023.3245672
23. J. Ren et al., “Latency Optimization in MEC Networks Using AI Techniques,” IEEE Transactions on Network and Service Management, 2023.
24. https://doi.org/10.1109/TNSM.2023.3256789
25. K. Zhang et al., “Energy-Latency Tradeoff in Edge Computing for IoT,” IEEE Internet of Things Journal, 2022. https://doi.org/10.1109/JIOT.2022.3147890
26. S. Barbarossa et al., “Communications and Control for Wireless Drone-Based Systems,” Proceedings of the IEEE, 2022. https://doi.org/10.1109/JPROC.2022.3145673
27. H. Guo et al., “Machine Learning-Based Task Offloading in IoT Systems,” Future Generation Computer Systems, 2023. https://doi.org/10.1016/j.future.2023.06.012
28. Z. Ning et al., “Joint Computation Offloading and Resource Allocation in Edge Computing,” IEEE Transactions on Vehicular Technology, 2022.
29. https://doi.org/10.1109/TVT.2022.3156784
30. X. Wang et al., “Dynamic Task Scheduling in Edge Computing for IoT,” Computer Networks, 2023. https://doi.org/10.1016/j.comnet.2023.109743
31. Y. He et al., “Multi-Agent Reinforcement Learning for Task Offloading in Edge Computing,” IEEE Transactions on Wireless Communications, 2025.
32. https://doi.org/10.1109/TWC.2025.3456789
33. J. Tang et al., “Semantic Segmentation-Based Data Offloading in IoT Systems,” Future Generation Computer Systems, 2025. https://doi.org/10.1016/j.future.2025.12.015
34. Shinde , S., Gadge , T., Dhaygude , V., Dhanrale , Y., & Rahane ,S.L. (2025). IoT-Based Smart Agriculture on the Cloud. International Journal of Recent Advances in Engineering and Technology, 14(1s), 257–260. https://doi.org/10.65521/intjournalrecadvengtech.v14i1s.287.
35. Ivailo Qudratullah. (2025). Artificial Intelligence Techniques for Efficient Energy Management in IoT-Enabled Large Buildings: Giant Trevally Optimizer (GTO) based Electric Vehicle Scheduling, Distributed Resource Integration, and Demand Response Strategies: Trends and Challenges. International Journal of Recent Advances in Engineering and Technology, 14(1), 359–366. https://doi.org/10.65521/intjournalrecadvengtech.v14i1.2562