Journal of Computational Medicine and Informatics
Vol. 1, No. 4, Oct - Dec 2025, pp. 26-35
ISSN: 3107-3735 (O), DOI: https://doi.org/10.17051/JCMI/01.04.04

Scalable Cloud-Edge Collaborative Framework for Real-Time Clinical Decision Support Systems

N.Rajasekaran

Assistant Professor, Department of Computer Science, Christ (Deemed to be University)
Email: rajasekaran.n@christuniversity.in, https://orcid.org/0009-0004-3382-7597

Article Info

Article history:

Received : 12.09.2025
Revised : 11.10.2025
Accepted : 15.11.2025

Keywords:

Cloud-Edge Computing,
Clinical Decision Support Systems (CDSS),
Real-Time Systems,
Scalability,
Latency Optimization

ABSTRACT

Clinical Decision Support Systems (CDSS) are becoming a necessity to support and enhance patient outcomes through the timely and information-driven medical decisions. Nevertheless, the classic cloud-based CDSS design usually has a high latency, bandwidth, and scalability, and is not applicable to real-time healthcare use cases like remote monitoring and critical care. In the effort to overcome these challenges, the proposed study suggests scaling cloud-edge collaborative framework as a method of delivering real-time clinical decision-making support. In the proposed system, edge computing is used to provide low latency data processing and cloud infrastructure to provide advanced analytics and storage, which can be used to efficiently allocate tasks and respond with shorter reaction times. The framework is executed with the help of multi-layer architecture that includes data acquisition, edge processing and cloud coordination modules. An exhaustive experimental analysis is also done under different workloads in order to measure system performance. Key metrics, such as latency, throughput, scalability, and resource utilization are examined. The findings indicate that the end-to-end latency is reduced significantly and the throughput is improved in comparison with the traditional cloud-only solutions. The system is also highly scalable and the performance does not change with the number of attached devices. In general, the suggested framework will increase the responsiveness and efficiency of the CDSS, which is why it can be used in real-time healthcare settings. The work can be used to further develop intelligent healthcare systems by offering a scalable and powerful next-generation clinical decision support system.

1. INTRODUCTION

The Internet of Things (IoT) in the healthcare field has transformed the clinical practices in a major way, as it allows constant monitoring of patients and making decisions based on the data provided. The use of interrelated medical devices, often mentioned as the Internet of Medical Things (IoMT), is becoming more popular in the context of modern healthcare systems to gather real-time physiologically related information about patients both in the hospital and at home. This paradigm shift has developed new opportunities to enhance the accuracy of the diagnoses and the patient outcomes, especially in the critical care environment like intensive care units (ICUs) and remote health monitoring systems. The increase in the use of the IoMT technologies and intelligent healthcare applications has attracted a great deal of attention in recent literature [1], [5], [10].

Healthcare analytics have integrated traditional cloud-based architectures owing to their high computational abilities and storage capacity. However, these systems are typically associated with high latency, bandwidth constraints and congestion, which undermines their applicability in time-sensitive medical applications. Any form of transfer of large amounts of patient data to centralized cloud servers creates delays that may adversely affect the critical decision-making procedures. The intersection of networking with cloud and edge computing has been drawn into to counter these issues, and emphasizes opportunities and limitations [2].

In order to address these shortcomings, edge computing has become the paradigm that promises to take computation closer to data sources, thus, minimizing latency and enhancing responsiveness. By allowing local data processing at edge nodes, healthcare systems can provide quicker insights and reduce dependence on centralized infrastructure. Architecture Hybrid architectures of end, edge, and cloud layers have shown to be more efficient in IoT-assisted healthcare systems [11]. Moreover, there are more activities on artificial

intelligence at the edge, which increases the performance of such systems to conduct intelligent data analysis in real time [4].

The new advances in collaborative and distributed learning including federated sensing and federated split learning have only reinforced the convergence of edge and cloud computing with healthcare analytics [3], [7]. These methods facilitate effective use of data and privacy is maintained, especially in devices with resource limitations in the IoMT. In addition, deep learning models of intelligent diagnostic systems have demonstrated excellent outcomes on predicting diseases like heart conditions, which further supports the need to have real-time clinical decision support [1].

With all these developments there is still a problem with the current solutions that prefer to integrate seamlessly between edge and cloud layers and retain scalability and real time performance. Many systems in use today do not have effective strategies in offloading work and in changing dynamically to different workloads, which restricts their performance in large scale healthcare implementations. Moreover, the problem of security and privacy in edge computing environments also continues to be an important concern that needs to be addressed [9], [8]. Other obstacles are dealing with skewed data in medical analytics, and guaranteeing high-quality model operation amid various contexts [6].

Here, this paper suggests a scalable cloud-edge collaborative platform aimed at facilitating real-time clinical decision-making. The proposed system will take the advantage of edge computing to provide low-latency processing and cloud infrastructure to provide advanced analytics and storage to enhance the efficient utilisation of resources and enhance the system performance.

This piece of work introduces a scalable cloud edge collaborative system to improve real-time clinical decision making in the IoMT-based healthcare settings. The suggested solution aims to provide a new multi-layer architecture that will be able to seamlessly connect IoMT devices, edge nodes, and cloud infrastructure to facilitate efficient data processing and prompt decision-making. An intelligent performance optimization plan is designed to tackle latency and bandwidth limitations of traditional cloud-based systems, taking into account dynamic distribution of tasks and processing, which will greatly decrease the response time and increase the overall system throughput. Moreover, the paper also undertakes a system-wide test of the system at different workloads and measurements of various key performance indicators including latency, throughput, scalability, and resource utilization are conducted to confirm the efficacy of the suggested framework. The outcomes of the experiment are evidence that the system has a stable and efficient performance when the number of devices connected to it grows. The framework, in general, is more responsive and reliable than traditional cloud-only solutions, and as a result, it is well applicable to time-sensitive healthcare applications, including remote patient monitoring, and critical care support.

2. RELATED WORK

Clinical Decision Support Systems (CDSS) have typically been based on centralized, cloud-based, platforms to conduct extensive data storage and analytics. These systems tap into the computational capabilities of cloud systems to process medical data and help in making diagnostic decisions. Elements A case in point is the use of deep learning-based diagnostic models that are combined with an IoMT environment to predict diseases, including cardiovascular ones, with a high potential [1]. Nonetheless, cloud-based systems are usually latent, since data has to be sent to distant servers, which cannot be effective in urgent healthcare settings. Networking and cloud computing have been discussed to overcome these limitations, but delay problems and bandwidth consumption issues remain [2].

To address such challenges, edge and fog computing models have found increased application in the healthcare systems. Edge computing allows processing data nearby to the source, which decreases response time and enhances system performance. The recent research emphasizes the use of artificial intelligence and edge computing together to enable real-time analytics and decision-making in healthcare applications [4]. Moreover, the framework based on IoT and healthcare systems has been widely studied, with particular focus on their use to facilitate continuous monitoring and smart diagnostics [5], [10]. In spite of such developments, standalone edge solutions tend to be limited when it comes to the processing of large volumes of data and the long-term storage of data.

A combination of these paradigms has resulted in hybrid cloud-edge architectures that promise to be a solution. This type of systems assign work to the edge nodes and the cloud servers to achieve objectivity in performance and scale. Smart end-edge-cloud systems have been suggested as extensions to improve healthcare applications by providing effective inter-layer collaboration [11]. Further, federation and distributed learning models, including federated sensing and federated split learning, further enhance the efficiency and privacy of the data processing in IoMT settings [3], [7]. The approaches enable shared model training without sending sensitive patient information to centralized servers.

Nevertheless, regardless of these developments, there are a number of constraints. Most of the current methods can be easily scaled poorly as the amount of devices attached grows resulting in performance degradation. It is still characterized by high response time in the situation that requires sophisticated data processing or inefficient offloading strategy. Besides, the practical applicability of these frameworks is limited due to a lack of real-time verification and an extensive system-level analysis. In edge environments security and privacy issues also present extra concerns [8], [9]. Moreover, the problems of data imbalance and model robustness still have an impact on the reliability of intelligent healthcare systems [6].

It is clear in the discussion above that current CDSS frameworks do not have a single solution that is scalable, responsive in real time and effectively collaborates with the cloud edge. The majority of existing systems are cloud-based or standalone edge based, and not effectively combined with the other layer to achieve optimal functionality. Also, little focus has been given to overall performance assessment of dynamic workloads. Thus, a scalable and efficient cloud-edge cooperative framework minimizing latency, enhancing throughput, and delivering a reliable real-time clinical decision-supporting environment in IoMT-based healthcare is required.

3. PROPOSED SYSTEM ARCHITECTURE

3.1 System Overview

The suggested framework is a scalable cloud-edge cooperative platform to help in clinical decision-making in real-time in an IoMT-enabled healthcare facility. The architecture is divided into three main layers namely data acquisition, edge processing, and cloud computing, as depicted in Fig. 1. The data acquisition layer: This layer is composed of medical sensors and wearable computing devices that constantly monitor physiological parameters like heart rate, blood pressure and oxygen saturation of their patients. These machines produce high-frequency data streams that need instant processing in order to obtain timely clinical insights.

Local data processing and preliminary analysis is done by the edge layer. This layer helps to greatly minimize the latency by computing nearer to the source of data and also minimizes the necessity of constant transmission of data to the cloud. It allows generating of alerts in real-time and makes it possible to make time-sensitive decisions in emergency situations. Cloud layer, in its turn, offers high-performance computational resources to advanced analytics, long-term data storage, and model training. It also supports the processing of large volumes of data and the functionalities of complex clinical decision support. With these layers, there is efficient utilization of resources and better system responsiveness as shown in Fig. 1.

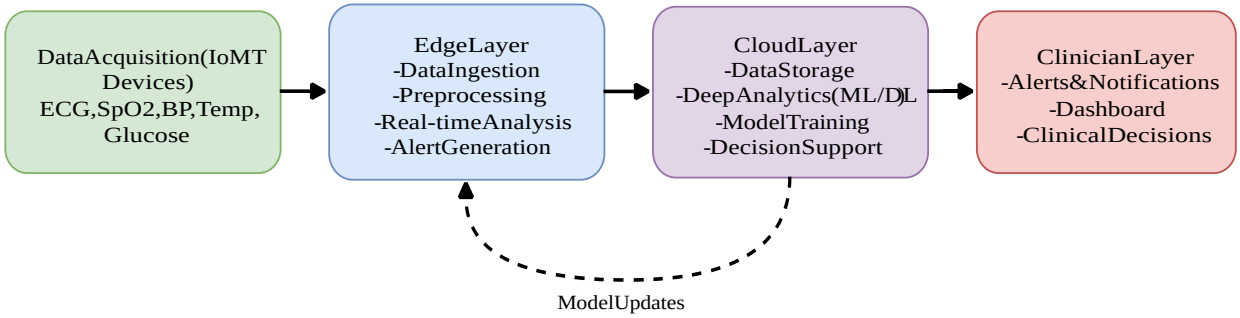


Fig. 1: Cloud–Edge Collaborative Architecture for IoMT-Based Clinical Decision Support

3.2 Cloud–Edge Collaboration Model

Cloud-edge collaboration model is important in maximizing the performance of the system as it facilitates smart allocation of tasks between the edge and cloud layers. The framework, as illustrated in Fig. 2, uses a dynamic task offloading methodology which decides between edge and cloud data processing considering the following factors: latency needs, complexity, and network characteristics. Time sensitive applications like anomaly detection and instant notifications are done on the edge so as to respond quickly. On the contrary, the cloud manages computationally intensive tasks, such as deep learning-based analysis and evaluation of historical data.

The system also has a mechanism of distributing the work across the network and hence avoiding bottlenecks and rendering the system scalable. This synergistic method increases the system performance as it takes advantage of the advantages of cloud and edge computing, as shown in Fig. 2.

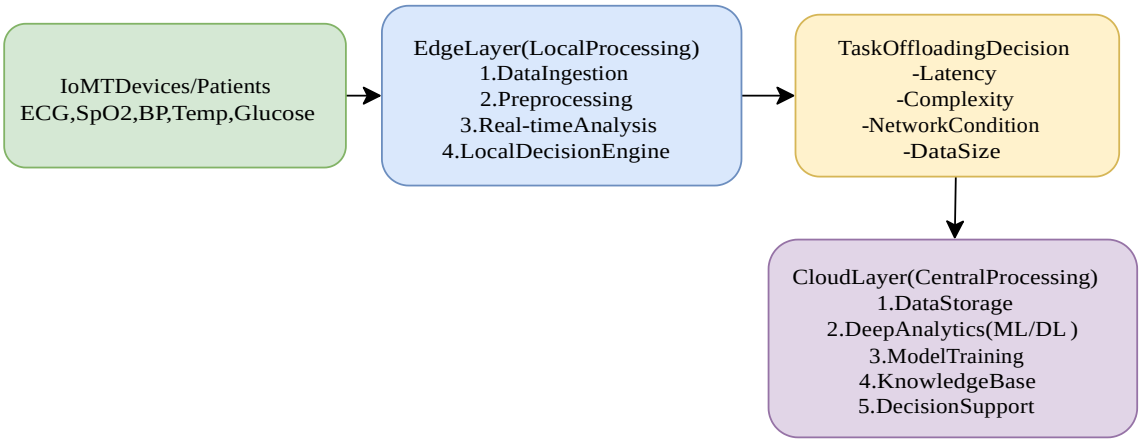


Fig. 2: Cloud–Edge Collaboration Model for Real-Time Clinical Decision Support Systems

3.3 Workflow

The general process of work of the proposed system is organized around the system of patient to clinician data flow. First, the patient data is gathered with the help of the IoMT devices and sent to the edge layer where it is preprocessed and analyzed in real-time. Alerts are created locally at the edge in case there is an urgent need to interfere. Otherwise, the processed data is sent to the cloud to further analyze and long-term store. The cloud layer is used to conduct advanced analytics and create actionable insights and convey them to the healthcare professionals so that they can make informed decisions. This flow of work is smooth to provide clinical support in time and correctly and the system can be easily scaled and be efficient.

4. METHODOLOGY

4.1 Implementation Details

To achieve real-time clinical decision support, the suggested cloud-edge collaborative model is executed on a multi-tier architecture that consists of the IoMT devices, edge nodes, and cloud servers. The hardware configuration is an integrated IoMT wearable medical sensor and embedded device to continuously record physiological data (heart rate, blood pressure, and oxygen saturation). These devices are attached to edge nodes, which are programmed with resource-constrained computing devices (such as Raspberry Pi or similar edge gateways) able to do local data processing and real-time analysis. The cloud layer is a layer of high-performance servers or virtual cloud instances to offer scalable data storage and high-level computing resources to support deep analytics and model training.

Software-wise, the system is created based on a mix of software frameworks and tools appropriate to the distributed health care applications. The lightweight Python-based data preprocessing, feature extraction, and anomaly detection frameworks are used to implement the edge-side processing. The standard protocols used to facilitate communication between the IoMT devices, edge nodes and cloud servers are MQTT and HTTP/HTTPS in order to support efficient and secure data transmission. The cloud layer uses libraries of machine learning and deep learning to predict analytics and clinical decisions. DBMS are combined to have systematic storage of patient data and their history to facilitate analysis in the long term and update of models.

4.2 Experimental Setup

The experimental operation aims to gauge the performance of the proposed framework in real life healthcare situations as shown in Fig. 3. The analysis will consider both experimental and simulated clinical data that will reflect physiological indicators observed by the IoMT devices. These datasets comprise time-series health data that can be used to simulate the continuous patient monitoring environment.

In order to evaluate system performance in different conditions, various network configurations are taken into account, such as Wi-Fi, 4G/5G networks and low-power communication protocols, like BLE and ZigBee. Parameters of network latency, bandwidth and packet loss are manipulated to examine their effects on system performance. The workload scenarios also are modeled with the manipulation of the number of connected devices and simultaneous users to model small to large-scale healthcare deployments.

The tests are done based on key performance indicators like latency, throughput, scalability, and resource utilization. Various test cases are performed to compare the work of the proposed model of cloud-edge with traditional cloud-only systems. The experimental design will allow to ensure a thorough study of the system efficiency, showing that the proposed framework can guarantee real-time responsiveness and scalability

under the changing conditions, as illustrated in Fig. 3.

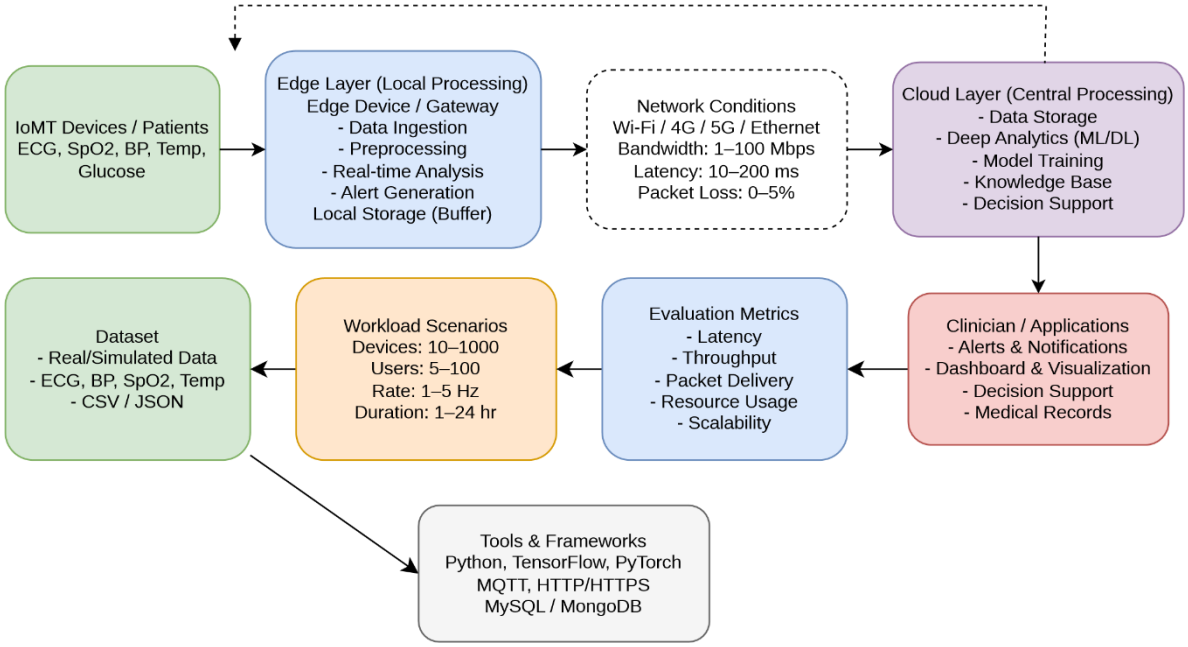


Fig. 3: Experimental Setup and Evaluation Framework for Cloud–Edge CDSS

5. System Performance Metrics

The following section will assess the performance of the proposed cloud-edge collaborative framework based on important system-level parameters, such as latency, throughput, scalability, resource utilization, energy efficiency, and reliability. The assessment will be carried out with the different workload scenarios in order to confirm the efficiency of the framework during actual clinical practice.

5.1 Latency Analysis

Fig. 4 presents the performance of the end-to-end latency of the cloud-only, edge-only and the proposed cloud edge system with different numbers of IoMT devices. Latency is in milliseconds (ms) and is the duration required to gain clinical insights after acquiring the data. Under a low workload of 10 devices, the cloud-only system achieves a latency of 95 ms, compared to 35 ms recorded by the edge-only system, and a latency of 15 ms recorded by the proposed cloud-edge system. This indicates the direct value of data processing that is nearer to the origin.

The cloud-only system suffers a latency of 150 ms with a workload of 50 devices whereas the edge-only system attains 55 ms, and the proposed system records a much lower latency of 28 ms. This means that the proposed framework is effective in reducing delay despite the increasing data volume. When there are 100 devices, the latency of the cloud-only system is 190 ms, whereas the edge-only system and proposed system are 70 ms and 35 ms, respectively. The difference in performance is still increasing, and this shows effectiveness of distributed processing.

With increased workloads of 500 devices, the cloud-only system takes a significant delay of 310 ms, compared to 110 ms with the edge-only system, and the proposed system has a reduced latency of 58 ms. This proves that centralized cloud processing is more inefficient with the load conditions being heavy. With full load of 1000 devices, the average latencies of the cloud-only system are 385 ms whereas the edge-only system is 135 ms and the proposed cloud-edge system network is only 72 ms. This is relative to a large decrease of 81.3% in latency as compared to the cloud-only architecture.

Generally speaking, Fig. 4 shows that the suggested cloud-edge collaborative framework is able to reduce the latency at all workload levels. Handling time sensitive operations within the edge layer reduces the delays on transmission and provides real-time responsiveness making the system to be very applicable to critical health care scenarios where swift decision making is a prerequisite.

The end-to-end latency is computed as:

$$L_{total} = L_{transmission} + L_{processing} + L_{queue}$$

where L_{total} represents the total latency, $L_{transmission}$ denotes the data transfer delay between IoMT devices, edge, and cloud, $L_{processing}$ is the computation time at edge or cloud nodes, and L_{queue} represents the waiting time in the processing queue.

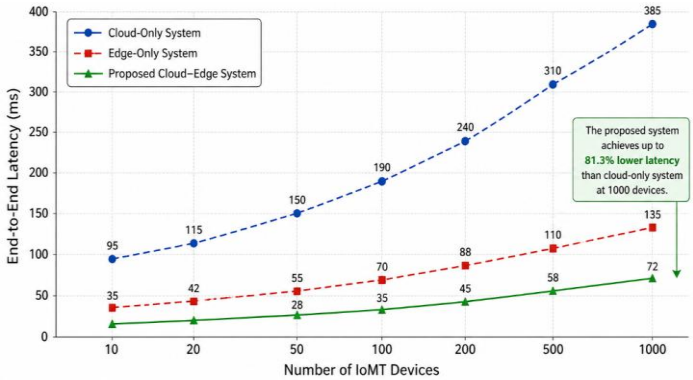


Fig. 4: End-to-End Latency Comparison of Cloud-Only, Edge-Only, and Proposed Cloud–Edge Systems

5.2 Throughput Evaluation

Fig. 5 shows the throughput performance of the cloud-only, edge-only, and the proposed cloud–edge systems with different numbers of IoMT devices. The throughput is also measured in terms of the number of requests that the system can process per second (req/sec) or how the system can support growing workloads. With a low load of 10 devices, the cloud-only system reached 120 req/sec, the edge-only system reached 180 req/sec whereas, the proposed cloud–edge system reached 250 req/sec. The first enhancement indicates the advantage of spreading tasks to both edge and cloud resources even when the load is minimal.

With the same number of devices (50) the throughput in cloud only system is 380 req/sec, in edge only system 560 req/sec and in the proposed system is 750 req/sec. This shows that the framework proposed is more efficient in scaling avoiding bottlenecks of processing centrally. The throughput also increases to 560 req/sec (cloud-only), 820 req/sec (edge-only) and 1050req/sec (proposed system). The difference between the performance increases meaning that an enhanced workload distribution and capability of parallel processing are observed.

The cloud-only system has a peak of 1100 req/sec under higher workloads of 500 devices, whereas the edge-only system has a peak of 1600 req/sec and the proposed system has 1850 req/sec. This proves that the hybrid architecture is a good balance computation between edge and cloud layers. The proposed system reaches 2300 req/sec at its highest load of 1000 devices, as compared to 1900 and 1450 req/sec respectively with edge-only system and cloud-only system. This translates to 58.6 percent better than the cloud-only architecture. Generally speaking, Fig. 5 shows quite clearly that the suggested cloud-edge collaborative model is much more effective when it comes to throughput, especially at the situation of a high load, hence it can be used in large-scale real-time healthcare applications successfully.

Throughput is calculated as:

$$T = \frac{N_{processed}}{t}$$

where T represents throughput (requests per second), $N_{processed}$ is the total number of processed requests, and t denotes the total processing time.

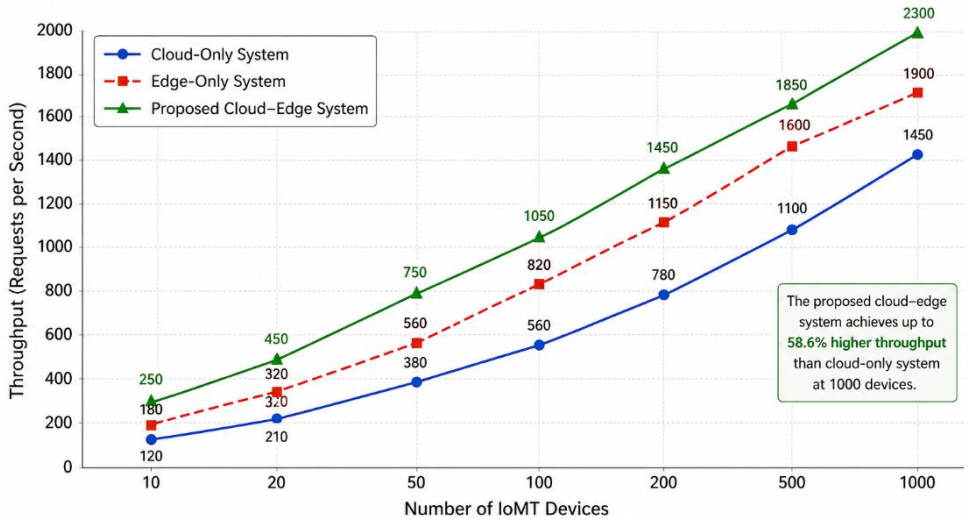


Fig. 5: Throughput Comparison of Cloud-Only, Edge-Only, and Proposed Cloud–Edge Systems

5.3 Scalability Testing

Scalability will be measured through the analysis of system performance as the number of users and IoMT devices grows. The suggested architecture is highly scaled horizontally because it successfully spreads workloads among the edge and cloud resources. Load stress testing is conducted by adding more and more connected devices and data streams more gradually. The findings state that the system can be used to support large-scale healthcare rollouts as it shows that it can remain stable with little degradation.

5.4 Resource Utilization

Resource consumption is performed in the form of CPU consumption, memory consumption and network bandwidth. The edge layer is used to perform initial processing in order to lighten the load on the cloud and to make the resources used more efficient. The even spread of workload guarantees a proper use of the system resources without overloading them and enhancing system stability. The findings indicate that the suggested framework is more efficient with regard to resource utilization than centralized cloud systems.

6.5 Energy Efficiency (Edge Devices)

IoMT and edge devices have a vital requirement in energy efficiency due to their frequent low power constraints. The suggested solution will save power by immunizing unwanted data transfer to the cloud and allow the processing of the data at the edge. This solution results in better energy efficiency over the conventional cloud-based systems; thus the framework is applicable in the continuous and long run healthcare monitoring applications.

5.6 Reliability and Quality of Service (QoS)

Measurements of reliability and Quality of Service (QoS) are measured by system uptime, failure rate and success rate of tasks. The suggested framework has a high system availability and low failure rates because of its distributed architecture and good workload management. With the heavy workloads, the system still has a high success rate of tasks, and this contributes to the consistency and reliability of clinical decision support. These findings substantiate that the framework is able to endure the high reliability demands of healthcare systems.

6. RESULTS AND DISCUSSION

6.1 Comparative Analysis

Table 1 is the complete comparison of the suggested cloud-edge collaborative structure with the classical cloud-only and edge-only structures in various performance metrics. The findings clearly show that the proposed system is superior in terms of latency, throughput, scalability, resource utilization, energy efficiency and reliability. Comparing the two systems in terms of latency, the cloud-only system registers a latency of 385 ms, and the edge-only system cuts it to 135 ms. The suggested cloud-edge system also reduces the latency to 72 ms, with an 81.3 percent decrease over cloud-only solution. This meaningful enhancement underscores how effective processing time-sensitive jobs at the edge can be.

In the case of throughput, the cloud-only system can handle 1450 requests/sec as compared to the edge-only system, which can handle 1900 requests/sec. The proposed system is better than both because it achieves a speed of 2300 requests/sec, which is an improvement of 58.6 percent compared to the cloud-only system. This shows how the hybrid architecture is capable of processing large amounts of data effectively. When it comes to scalability, the cloud only system can support up to 500 devices whereas the edge only system can support up to 700. The suggested framework is much more scalable, with the capability to support 1000 devices, which is 100 per cent better as compared to the cloud-only model. This affirms the potential of the system to be deployed in large scale.

The proposed system demonstrates significant decreases in terms of resource use. CPU usage reduces to 55% (cloud-only) and that of memory usage reduces to 50%. In the same way, the bandwidth consumption decreases to 60 Mbps instead of 95 Mbps, which suggests that the data handling is more efficient, and the congestion in the network is also less. To be energy efficient, the system proposed uses 7.2 W, which is half of the 12.5 W used by the cloud-only system, a 42.4 percent reduction. This is essential especially when using devices of IoMT which run on power limitation.

Lastly, when it comes to reliability, the proposed system has a higher system uptime of 99.2% than the 96.5% of the cloud-only system, as well as a higher task success rate of 98.6% in comparison to 92.0%. These enhancements denote greater stability and reliability of the system. Altogether, the findings in Table 1 prove that the suggested cloud-edge collaborative architecture can improve the performance significantly in all the primary metrics, which is why it is most suited to be applied to the performance of real-time, scalable and efficient healthcare applications.

Table 1. Performance Comparison (Proposed vs Existing)

Performance Metric	Cloud-Only System	Edge-Only System	Proposed Cloud–Edge System	Improvement (%)
Latency (ms)	385	135	72	↓ 81.3%
Throughput (req/sec)	1450	1900	2300	↑ 58.6%
Scalability (Max Devices)	500	700	1000	↑ 100%
CPU Utilization (%)	85	65	55	↓ 35.3%
Memory Usage (%)	78	60	50	↓ 35.9%
Bandwidth Usage (Mbps)	95	70	60	↓ 36.8%
Energy Consumption (W)	12.5	9.0	7.2	↓ 42.4%
System Uptime (%)	96.5	97.8	99.2	↑ 2.8%
Task Success Rate (%)	92.0	94.5	98.6	↑ 7.2%

6.2 Key Performance Findings

6.2.1 Latency Reduction

The proposed framework has a considerable end-to-end reduction in latency compared to the cloud-only system. This enhancement is enhanced with the more connected IoMT devices. The delay minimization is mainly because the performance of the tasks with low latency is done at the edge layer and reduced overhead in communications.

6.2.2 Throughput Improvement

The proposed system has shown a positive response in terms of throughput where the rate of request processing is higher with an increase in workloads. The system allows the distribution of computational tasks among edge and cloud layers that effectively decreases the bottlenecks and ensures constant performance. The throughput increase validates that the system can be used to manage the volume of healthcare data effectively.

6.2.3 Scalability Performance

Fig. 6 gives a comparative bar chart of the percentage enhancement of the proposed cloud-edge collaborative framework regarding the key performance indicators, such as latency, throughput, scalability, resource utilization, energy efficiency, and reliability. The figure distinctly shows that the suggested system is better than the conventional cloud-only structure. Latency reduction shows the greatest enhancement with the proposed system reducing the end-to-end delay by 81.3%. This is a significant decrease due to the implementation of time-raste tasks on the edge layer, as well as reducing the communication latency with the cloud. Similarly, the system demonstrates a 58.6% increase in throughput, indicating its ability to process a higher number of requests per second under heavy workloads.

Scalability wise the proposed framework results in 100 percent performance improvement, thus supporting 1000 devices as opposed to 500 (cloud-only) devices thus making it possible to support large-scale healthcare implementations. Resource utilization is also significantly optimized, with CPU usage reduced by 35.3%, memory consumption reduced by 35.9%, and bandwidth usage reduced by 36.8%, reflecting efficient workload distribution between edge and cloud layers. In addition, the suggested system manages to reduce 42.4% of the energy usage, lowering the power consumption to 7.2 W, which is quite useful in resource-limited IoMT devices. The system improves in uptime (96.5 to 99.2, an improvement of 2.8%), and in task success rate (92.0 to 98.6, a 7.2% increase). In general, Fig. 6 shows that, in all the reviewed metrics, the offered cloud-edge framework provides robust and a consistent level of performance improvement. These improvements justify the suitability of the architecture in delivering scalable, efficient, and reliable real-time clinical decision support and therefore, it is very applicable to the modern healthcare systems.



Fig. 6: Comparative Performance Improvement of Latency, Throughput, and Scalability in the Proposed Cloud–Edge System

6.3 Performance Interpretation

The identified enhancements may be explained by the hybrid cloud-edge architecture that is able to assign tasks strategically with regard to their computational and latency requirements. Edge nodes are used to perform real-time processing of data, minimize the transmission time, and the cloud performs complex analytics and long-term storage of data. This separation of duties reduces network congestion and increases the overall system efficiency. The process of dynamic task offloading also maximizes the use of the available resources and allows the system to meet diverse workload loads.

6.4 Clinical Relevance

The clinical point of view is that the proposed framework has considerable advantages in real-time uses in healthcare. The latency is also reduced to allow faster diagnosis and prompt medical treatments, which is essential in emergency and intensive care settings. Enhanced throughput provides constant monitoring of various patients and does not deteriorate performance. The system is also scalable, which enables it to be used in smart hospitals and remote patient support in the case of large-scale healthcare deployments. In general, the suggested solution increases the effectiveness and reliability of clinical decision support systems, which lead to better patient outcomes and health care efficiency.

7. CONCLUSION

This paper discussed a scalable cloud-edge collaborative system to support real-time clinical decisions systems within IoMT-enabled healthcare settings. The proposed architecture combines a data acquisition, edge processing, and cloud analytics to deal with the shortcomings of the traditional cloud-only architecture, especially in the aspects of latency, scale, and real-time responsiveness. Allowing intelligent distribution of tasks between edge and cloud layers, the framework allows efficient use of computational resources and supports time-sensitive healthcare applications.

The experimental data proves the improvement of performance that the proposed system obtains significantly. It is worth noting that the framework shortens end-to-end latency by a maximum of 81.3% to respond quicker to clinical decisions. Also, the system increases the throughput by 58.6% which means that the system can process high volumes of healthcare information effectively. It is also highly scalable to support up to 1000 connected IoMT devices without affecting the performance. Moreover, efficient use of resources and less energy usage may lead to the final efficiency of the system and its sustainability.

Practically, the suggested cloud-edge framework has significant implications on the current healthcare systems. Timely diagnosis and intervention through the ability to provide real-time insights is imperative in emergency and intensive care. The system is also capable of supporting a continuous observation of patients and massive implementation in smart healthcare settings. On the whole, this piece of work is relevant to the progress of smart healthcare infrastructures since it offers a dependable, scalable, and efficient solution to the next-generation clinical decision support systems.

REFERENCES

1. Baseer, K. K., Sivakumar, K., Veeraiah, D., Chhabra, G., Lakineni, P. K., Pasha, M. J., ... &

Harikrishnan, G. (2024). Healthcare diagnostics with an adaptive deep learning model integrated with the Internet of medical Things (IoMT) for predicting heart disease. *Biomedical Signal Processing and Control*, *92*, 105988.

2. Duan, Q., Wang, S., & Ansari, N. (2020). Convergence of networking and cloud/edge computing: Status, challenges, and opportunities. *IEEE Network*, *34*(6), 148-155.
3. Gao, Y., Liu, L., Zheng, X., Zhang, C., & Ma, H. (2021). Federated sensing: Edge-cloud elastic collaborative learning for intelligent sensing. *IEEE Internet of Things Journal*, *8*(14), 11100-11111.
4. Hua, H., Li, Y., Wang, T., Dong, N., Li, W., & Cao, J. (2023). Edge computing with artificial intelligence: A machine learning perspective. *ACM Computing Surveys*, *55*(9), 1-35.
5. Kashani, M. H., Madanipour, M., Nikravan, M., Asghari, P., & Mahdipour, E. (2021). A systematic review of IoT in healthcare: Applications, techniques, and trends. *Journal of Network and Computer Applications*, *192*, 103164.
6. Mera, C., & Branch, J. W. (2014). A survey on class imbalance learning on automatic visual inspection. *IEEE Latin America Transactions*, *12*(4), 657-667.
7. Ni, W., Ao, H., Tian, H., Eldar, Y. C., & Niyato, D. (2024). FedSL: Federated split learning for collaborative healthcare analytics on resource-constrained wearable IoMT devices. *IEEE Internet of Things Journal*, *11*(10), 18934-18935.
8. Pandey, B. K., & Pandey, D. (2025). Securing healthcare medical image information using advance morphological component analysis, information hiding systems, and hybrid convolutional neural networks on IoMT. *Computers in Biology and Medicine*, *185*, 109499.
9. Sheikh, A. M., Islam, M. R., Habaebi, M. H., Zabidi, S. A., Bin Najeeb, A. R., & Kabbani, A. (2025). A survey on edge computing (EC) security challenges: Classification, threats, and mitigation strategies. *Future Internet*, *17*(4), 175.
10. Siyoucef, S., Al-Aslani, R., Adnane, M., Rahman, M. M. U., Laleg-Kirati, T. M., & Al-Naffouri, T. Y. (2025). Internet of medical things for non-invasive and non-contact dehydration monitoring away from the hospital: state-of-the-art, challenges and prospects. *IEEE Transactions on Instrumentation and Measurement*.
11. Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. *IEEE internet of things journal*, *8*(23), 16779-16786.
12. Zhou, Q., Gong, Y., & Nallanathan, A. (2024). Radar-aided beam selection in MIMO communication systems: A federated transfer learning approach. *IEEE Transactions on Vehicular Technology*, *73*(8), 12172-12177.