INTELLITHINGS
Artificial Intelligence | Internet of Things

Vol. 1 No. 1, 2024 p-ISSN: xxxx-xxxx e-ISSN: xxxx-xxxx https://doi.org/xxxxxxxxx

Adaptive Load Balancing Model on Edge Gateways to Support IoT Scalability in 5G Networks

Ata Amrullah

Department of Informatics, Darul Ulum Islamic University, East Java, Indonesia

Received: 2 December 2023
Accepted: 9 January 2024
Published: 29 February 2024

Keywords:
Load Balancing;
Edge Computing;
IoT Scalability;
Artificial Intelligence;
5G Networks.

Corresponding author:
Ata Amrullah
ata@unisda.ac.id

Abstract

The rapid proliferation of Internet of Things (IoT) devices, particularly within the context of smart cities, industrial automation, and connected vehicles, poses significant challenges to network scalability and real-time data processing. With the advent of 5G networks, promising ultra-low latency and massive connectivity, the role of edge computing and specifically edge gateways becomes critical. However, the dynamic and heterogeneous nature of IoT traffic, coupled with varying computational demands, can lead to uneven resource utilization and performance bottlenecks on edge gateways. This paper proposes an adaptive load balancing model designed to optimize resource distribution and enhance IoT scalability in 5G networks. We explore how artificial intelligence (AI) and machine learning (ML) techniques can be leveraged for real-time traffic prediction, dynamic task offloading, and intelligent resource allocation across multiple edge gateways. The proposed model aims to minimize latency, maximize throughput, and ensure high availability for diverse IoT applications. We discuss the architectural components, key adaptive mechanisms, and the integration with 5G network capabilities, alongside outlining persistent challenges and promising future research directions to build more resilient and efficient IoT ecosystems.

1. Introduction

The Internet of Things (IoT) paradigm is rapidly transforming various sectors, enabling unprecedented levels of connectivity and data generation from an ever-increasing number of heterogeneous devices [1]. From smart homes and healthcare to industrial automation and intelligent transportation systems, IoT deployments are becoming central to modern infrastructure. This proliferation, however, introduces significant challenges, primarily concerning the scalability of network infrastructure and the ability to process vast streams of data in real-time [2]. Traditional cloud-centric architectures, while offering centralized processing power, are inherently limited by latency, bandwidth constraints, and potential single points of failure, making them less suitable for time-critical IoT applications.

The emergence of Fifth Generation (5G) mobile networks marks a pivotal shift in addressing these limitations. 5G technology, with its promises of enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC), provides the foundational connectivity necessary for advanced IoT applications [3]. Specifically, 5G's integration with Multi-access Edge Computing (MEC) allows computational resources to be deployed at the network edge, closer to IoT devices. This synergy between 5G and edge computing (Edge-IoT) is crucial for meeting the demanding QoS requirements of real-time IoT services, such as autonomous vehicles, critical infrastructure monitoring, and augmented reality applications [4].

Within the Edge-IoT architecture, edge gateways play a vital role. These gateways act as intermediaries, aggregating data from numerous IoT devices, performing initial processing, and intelligently routing traffic either to other edge nodes or to the cloud [5]. They bridge the gap between resource-constrained IoT devices and powerful, but distant, cloud data centers. However, as the number of connected IoT devices and the complexity of their data streams grow, edge gateways can quickly become bottlenecks. Static or simplistic load balancing strategies are insufficient to handle the highly dynamic, unpredictable, and often bursty nature of IoT traffic and the varying computational loads it imposes [6]. An overloaded edge gateway can lead to increased latency, dropped packets, and degradation of service for critical IoT applications, thus undermining the promises of 5G and edge computing.

Therefore, an adaptive load balancing model on edge gateways is indispensable for supporting IoT scalability in 5G networks. Such a model must be capable of real-time monitoring of network conditions and gateway resources, intelligently predicting future loads, and dynamically distributing incoming IoT tasks or data streams to optimize overall system performance. This paper proposes a comprehensive adaptive load balancing model that leverages artificial intelligence (AI) and machine learning (ML) techniques to address these challenges.

Our main contributions are:

- To highlight the critical role of adaptive load balancing for IoT scalability in 5G-enabled edge computing environments;
- To propose an architectural model for adaptive load balancing on edge gateways, integrating AI/ML capabilities;
- To detail the key mechanisms of adaptation, including real-time monitoring, predictive analytics, and dynamic decision-making;
- To discuss the integration points with 5G network features like MEC for enhanced performance;
- To identify persistent challenges and outline promising future research directions in this vital area.

The remainder of this paper is organized as follows: Section 2 provides background information and reviews related work. Section 3 presents the proposed adaptive load balancing model. Section 4 discusses its performance considerations and optimization challenges. Finally, Section 5 concludes the paper and suggests future research avenues.

2. Background and Related Work

2.1. IoT Scalability Challenges

The proliferation of IoT devices brings forth immense scalability challenges. The sheer volume of connected devices, projected to reach tens of billions, demands network infrastructures capable of handling massive concurrent connections and diverse data types [7]. Beyond sheer numbers, heterogeneity in device capabilities (e.g., computational power, energy constraints), communication protocols (e.g., MQTT, CoAP, LoRaWAN), and application requirements (e.g., real-time control vs. periodic sensing) adds significant complexity [1]. Managing these diverse demands while ensuring Quality of Service (QoS) and low latency is a major hurdle for traditional network architectures.

2.2. The Role of 5G in IoT Evolution

5G networks are specifically designed to address the limitations of previous generations, offering transformative capabilities critical for large-scale IoT deployments:

a. **Enhanced Mobile Broadband (eMBB)**: Provides high data rates (up to 10 Gbps), essential for data-intensive IoT applications like high-definition video surveillance.

b. **Ultra-Reliable Low-Latency Communication (URLLC)**: Achieves latencies as low as 1 ms and high reliability, crucial for mission-critical IoT applications such as autonomous driving, remote surgery, and industrial automation [8].

c. **Massive Machine-Type Communication (mMTC)**: Supports connectivity for up to 1 million devices per square kilometer, enabling the widespread deployment of sensors and smart devices in smart cities and other large-scale environments [9].

d. **Network Slicing**: Allows creation of multiple virtual networks over a common physical infrastructure, each tailored to specific service requirements (e.g., a slice for URLLC IoT devices, another for eMBB devices), facilitating flexible resource allocation and QoS guarantees [10].

2.3. Edge Gateways in 5G-IoT Architectures

Edge gateways serve as critical intermediate nodes in the 5G-IoT ecosystem, bridging the gap between IoT devices and the core network or cloud. They are deployed at the network edge, often as Multi-access Edge Computing (MEC) servers, close to the data sources [4]. Their primary functions include:

a. Data Aggregation and Pre-processing: Collecting data from numerous heterogeneous IoT devices, filtering out noise, and aggregating relevant information.

b. Local Computation and AI Inference: Running AI models to perform real-time analytics, anomaly detection, or decision-making close to the data source, reducing reliance on cloud round-trips.

c. Protocol Translation: Interconnecting devices using different communication protocols.

d. Security Enforcement: Acting as a first line of defense for IoT traffic.

e. Load Distribution: Distributing incoming requests or tasks among available processing units or other edge nodes.

2.4. Existing Load Balancing Techniques

Traditional load balancing techniques, commonly used in data centers, include:

a. Static Methods: Such as Round Robin (distributing requests sequentially) or Weighted Round Robin (assigning weights to servers based on capacity). These methods are simple but fail to adapt to dynamic changes in load or server health [11].

b. Dynamic Methods: Such as Least Connections (directing new requests to the server with the fewest active connections) or Least Response Time (routing to the server responding fastest). While better, these often rely on instantaneous metrics and may not accurately predict future load or account for complex task types [11].

In the context of IoT and edge computing, recent research has started incorporating AI/ML:

a. [12] proposed an agent-based reinforcement learning (RL) approach for task offloading in MEC environments, demonstrating improved latency compared to static methods.

b. [13] utilized deep learning (DL) models to predict network traffic for proactive resource allocation in edge networks.

c. [14] explored a federated learning framework for distributed load balancing, where edge nodes collaboratively learn optimal policies without sharing raw resource data.

While these studies highlight the potential of AI/ML, a holistic adaptive load balancing model specifically designed for the unique dynamics and scalability demands of 5G-enabled IoT environments, especially focusing on edge gateways as the primary bottleneck, requires further comprehensive exploration.

3. Proposed Adaptive Load Balancing Model

To address the inherent challenges of IoT scalability in 5G networks, particularly the dynamic load on edge gateways, we propose an adaptive load balancing model. This model leverages real-time monitoring, predictive analytics through AI/ML, and intelligent decision-making to optimize resource utilization and enhance overall system performance.

3.1. Model Architecture Overview

The proposed architecture extends the typical 5G-IoT setup with specialized components for adaptive load balancing on edge gateways, as illustrated conceptually in Figure 1.

- IoT Devices Layer: Consists of diverse IoT devices generating data and service requests.
- Edge Gateways Layer: Multiple edge gateways, each equipped with:

  - Load Balancer Module: The core intelligence unit for adaptive load balancing.
  - Resource Monitor: Continuously collects real-time metrics (CPU usage, memory, network I/O, queue length, latency) of the local gateway and potentially neighboring gateways
  - AI/ML Prediction Engine: Utilizes historical data and real-time monitored metrics to predict future load patterns and resource availability
  - Decision Maker: Based on predictions and current system state, determines the optimal routing or offloading strategy for incoming IoT requests/tasks
  - Task/Traffic Router: Implements the decision from the Decision Maker, directing traffic to the optimal processing unit (local processing, other edge gateway, or cloud)
- 5G Core Network: Provides the underlying high-speed, low-latency connectivity, supporting network slicing and MEC functionalities.
- Cloud Layer: For aggregated data storage, long-term analytics, global AI model training, and handling tasks that are not time-critical or require extensive computational resources.

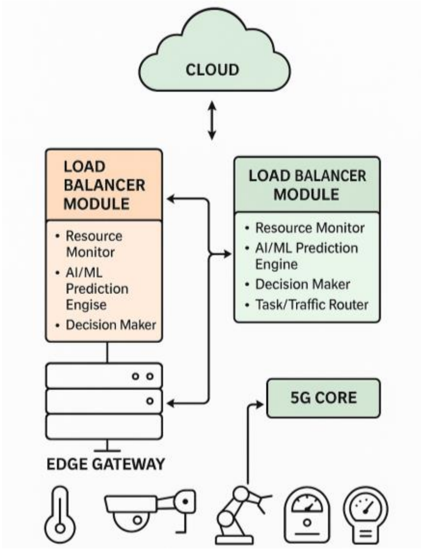

Figure 1. 5G-IoT setup with load balancing

3.2. Key Adaptive Mechanisms

The model's adaptiveness is achieved through a continuous feedback loop and intelligent processing:

a. Real-time Resource and Traffic Monitoring:
- Each edge gateway's Resource Monitor constantly gathers fine-grained data on its own operational status (e.g., CPU load, RAM usage, network bandwidth, processing queue size, latency to connected IoT devices) [15].
- Information exchange mechanisms (e.g., lightweight heartbeats or shared state via a distributed ledger) allow gateways to be aware of the real-time load and capacity of their neighboring edge gateways [16].
- Traffic characteristics (e.g., packet rate, data volume, application type, urgency) are also monitored at the ingress point.

b. AI/ML-Powered Prediction Engine
- The Prediction Engine is the brain of the adaptive model. It takes historical and real-time monitoring data as input to forecast future load trends and resource availability.
- Reinforcement Learning (RL): Q-learning or Deep Q-Networks (DQN) can be used where the load balancer agent learns optimal routing policies by interacting with the dynamic network environment. Rewards can be defined based on minimizing latency, maximizing throughput, or balancing resource utilization [17].
- Recurrent Neural Networks (RNNs) / Long Short-Term Memory (LSTM): Excellent for time-series data prediction, LSTMs can analyze historical traffic patterns and resource usage to anticipate upcoming load fluctuations [18]. This proactive approach allows the Decision Maker to prepare for anticipated spikes.
- Federated Learning: In a multi-gateway deployment, FL can be applied for collaborative learning of traffic patterns across different edge domains without centralizing sensitive operational data. Each gateway trains a local prediction model, and only model updates are exchanged, maintaining privacy and distributed intelligence [14].

c. Dynamic Decision Maker

- Based on the current state from the Resource Monitor and future predictions from the AI/ML Prediction Engine, the Decision Maker dynamically determines the optimal load balancing strategy for each incoming IoT request or data stream.
- Decision criteria can be configured based on several policies. First, to minimize latency, routing can be prioritized to the least loaded gateway or through local processing, especially for URLLC (Ultra-Reliable Low-Latency Communication) applications. Second, to maximize throughput, high-volume eMBB (enhanced Mobile Broadband) data streams can be distributed across the available bandwidths. Third, to balance resource utilization, the system should ensure that no single gateway is either over- or under-utilized, thereby promoting efficient energy consumption and preventing bottlenecks [3]. Lastly, to support application-specific Quality of Service (QoS), critical IoT traffic—such as from autonomous vehicles—can be routed to a gateway with guaranteed resources, potentially leveraging 5G network slicing [4].
- Actions that can be triggered by the Decision Maker include several response strategies. One option is local processing, where the request is handled directly on the current edge gateway. Another is inter-gateway offloading, which involves redirecting the request to a neighboring edge gateway that has a lighter load. For tasks that are not time-critical but require significant computational resources, cloud offloading can be employed. Additionally, the system can perform connection migration, dynamically transferring existing IoT device connections to another gateway if there is a significant change in load conditions.

3.3. Integration with 5G Network Capabilities

The proposed model inherently benefits from and integrates seamlessly with key 5G features. First, MEC (Multi-access Edge Computing) infrastructure is leveraged by deploying edge gateways as MEC servers, which provides the necessary computational and networking proximity to IoT devices. Second, the Decision Maker can utilize 5G network slicing capabilities to allocate dedicated network resources—such as bandwidth and guaranteed latency—to critical IoT applications, thereby optimizing their performance even under conditions of overall network congestion [4]. Third, with support for URLLC (Ultra-Reliable Low-Latency Communication) and mMTC (massive Machine Type Communications), the adaptive load balancing mechanism ensures that the edge infrastructure can fulfill 5G's promises of ultra-low latency and massive device connectivity by managing resource contention intelligently. By orchestrating these adaptive mechanisms, the proposed model elevates edge gateways from static intermediaries to intelligent, dynamic resource managers, ensuring optimal performance and true scalability for a wide range of IoT applications in the 5G era.

4. Performance Evaluation and Optimization Considerations

The effectiveness of the proposed adaptive load balancing model hinges on its ability to demonstrably improve key performance metrics in a dynamic 5G-IoT environment. This section outlines the critical evaluation parameters and discusses specific optimization challenges inherent in such a system.

4.1. Key Performance Metrics

To properly assess the model's effectiveness, several key performance metrics need to be considered. One of the most important is end-to-end latency — the time it takes for a request from an IoT device to receive a response — which is especially critical for real-time applications. Throughput is also vital, as it reflects how many requests or data packets the system can handle over time, indicating its capacity to manage large-scale traffic. Another important factor is how efficiently resources like CPU, memory, and network bandwidth are used across edge gateways; ideally, the load should be well-balanced to avoid overloading some nodes while others sit idle. Energy consumption matters too, particularly for smart city applications where sustainability is a priority — smart load balancing can help reduce unnecessary energy use by directing traffic to more efficient nodes. The model's scalability is also key, meaning it should perform well even as the number of connected devices and traffic increases, including being able to easily add or remove edge gateways as needed. Reliability and fault tolerance are just as important, ensuring the system can keep running smoothly even if one or more gateways fail. Finally, there's decision overhead — the extra work the system has to do to monitor performance and make load balancing decisions. This overhead needs to stay low enough that it doesn't cancel out the benefits the model provides.

4.2. Simulation and Real-world Deployment Considerations

When evaluating a complex model like this, both simulation and real-world testing play important roles. Simulation environments—such as NS-3, Mininet-WiFi, or custom-built simulators—are commonly used to replicate 5G networks, IoT device behavior, and interactions with edge gateways. These tools make it possible to experiment in a controlled setting, adjusting factors like traffic load and network conditions. However, for a more realistic assessment, deploying the model on a small-scale testbed with actual IoT devices and edge servers can reveal practical challenges, including hardware limitations and unpredictable environmental influences. Additionally, having access to diverse IoT traffic datasets—whether synthetic or collected from real-world sources—is crucial for training and evaluating the AI/ML prediction engine. These datasets should capture various device types, traffic patterns (such as bursty or periodic), and differing application demands to ensure the model performs well across a wide range of scenarios.

4.3. Optimization Challenges for the Model

While the proposed adaptive load balancing model offers significant advantages, it also comes with several optimization challenges. One major issue is the complexity of AI/ML models, which can cause inference delays when deployed on resource-limited edge gateways. Techniques like model pruning, quantization, or knowledge distillation are essential to reduce computational overhead and ensure real-time decision-making. Another challenge lies in the accuracy of load prediction—unpredictable traffic patterns or rare events can easily lead to incorrect forecasts and inefficient load distribution, so the model needs to be robust against noisy or outlier data. Additionally, adapting decision-making policies in dynamic environments is complex; the system must constantly balance conflicting goals, such as minimizing latency versus maximizing throughput, potentially using methods like meta-learning or multi-objective optimization. Keeping all edge gateways synchronized with the latest system state adds further strain, as frequent communication can introduce significant bandwidth usage and latency. Security is also a major concern, given that the load balancer acts as a central control point—protecting it from attacks like data poisoning or denial-of-service is critical, possibly through solutions like blockchain or trusted execution environments. Moreover, the system must account for the heterogeneity of edge resources, as gateways often differ in processing power, network quality, or energy constraints. Lastly, integrating the model with 5G network slicing introduces interoperability challenges, requiring standardized interfaces and advanced orchestration to ensure seamless coordination between application-level decisions and network-level resource management.

Addressing these challenges requires a careful balance between computational sophistication and practical deployment constraints. Future research should focus on designing lightweight, resilient, and secure AI/ML models that can perform effectively on edge devices with limited resources. Equally important is the development of intelligent orchestration frameworks capable of managing the dynamic and distributed nature of 5G-IoT environments, ensuring seamless integration, adaptability, and end-to-end system efficiency.

5. Challenges And Future Directions

The development and deployment of an adaptive load balancing model for edge gateways in 5G-IoT networks, while highly promising, is confronted by a complex set of challenges that not only test current technological capabilities but also highlight key areas for future research. These challenges span from computational constraints and prediction accuracy to security, interoperability, and dynamic policy adaptation—each requiring innovative solutions to ensure the model's effectiveness in real-world scenarios. As such, addressing them is essential to unlocking the full potential of intelligent, resilient, and scalable edge computing in next-generation IoT ecosystems.

5.1. Dynamic Environment Complexity and Predictability:

The constantly changing landscape of IoT traffic and fluctuating 5G network conditions—such as variable signal strength, congestion, and frequent handovers—poses a major challenge for accurate, long-term load prediction. These dynamic factors make it difficult for any model to maintain consistent performance over time. To address this, future research should explore hybrid prediction models that merge real-time data analysis with long-term historical patterns, leveraging advanced AI techniques like transformer-based architectures for time-series forecasting. Additionally, the development of robust anomaly detection mechanisms is essential to enable the system to quickly recognize and respond to unexpected spikes or drops in traffic that deviate from normal behavior, ensuring adaptive and resilient load balancing in unpredictable environments.

5.2. Resource Heterogeneity and Federation:

In smart city environments, edge gateways often differ widely in terms of processing power, memory, storage capacity, and network capabilities. This heterogeneity becomes even more challenging when these resources are distributed across different administrative domains—such as government agencies, private companies, or telecom providers—each with its own management policies and operational priorities. Balancing loads effectively in such a fragmented ecosystem requires innovative approaches. Future research should focus on federated resource orchestration, where frameworks are developed to enable collaborative, cross-domain resource sharing and load balancing. Integrating technologies like blockchain can help establish trust, transparency, and secure coordination among competing stakeholders. Additionally, resource virtualization and abstraction techniques are needed to create a consistent, unified view of diverse edge resources, allowing the load balancing model to make informed and simplified decisions without being burdened by underlying complexity.

5.3. AI Model Robustness, Explainability, and Continual Learning:

In critical IoT applications, the trustworthiness and reliability of AI-driven load balancing decisions are essential. One major concern is robustness against adversarial attacks, where malicious inputs or data poisoning could manipulate traffic predictions or routing decisions, potentially destabilizing the system. Ensuring AI models can withstand such threats is crucial for maintaining secure operations. Additionally, there's a growing need for explainable AI (XAI) at the edge—lightweight techniques that can provide clear, understandable justifications for decisions made by the AI, directly on resource-constrained gateways. This transparency supports debugging, auditing, and building stakeholder confidence in automated decision-making [19]. Furthermore, as IoT environments evolve rapidly, AI models must support continual learning—the ability to adapt incrementally to new device types, traffic patterns, and changing network conditions without needing to be retrained from scratch or heavily dependent on manual updates. Future research must prioritize these areas to ensure AI systems remain resilient, transparent, and adaptive in real-world deployments.

5.4. Security and Privacy of Load Balancing Decisions:

As a critical control plane element, the load balancing module must be highly secure since any compromise could result in denial-of-service attacks, data breaches, or malicious manipulation of traffic flows. To enhance trust and transparency, future research should explore blockchain-based solutions that use distributed ledger technology to record and verify load balancing decisions, ensuring they are immutable and easily auditable. Additionally, implementing privacy-preserving techniques such as homomorphic encryption or secure multi-party computation can protect sensitive resource usage data shared across multiple gateways, allowing collaborative decision-making without exposing confidential information. These approaches are vital to safeguarding both the integrity and privacy of the load balancing process in distributed edge environments.

5.5. Integration with Advanced 5G Features:

Although 5G technology provides powerful capabilities like network slicing, effectively leveraging these features through an adaptive load balancer requires further exploration. One key area is dynamic network slice allocation, which involves automatically requesting and releasing network slices based on real-time and predicted IoT traffic demands, ensuring optimal quality of service (QoS) tailored to different application needs. Another promising direction is the integration with intent-based networking (IBN), where high-level policy goals—such as guaranteeing latency below 10 milliseconds for critical infrastructure sensors—are automatically translated into precise load balancing and network configuration actions. This seamless coordination between application requirements and network capabilities can greatly enhance the efficiency and responsiveness of 5G-IoT systems.

5.6. Standardization and Interoperability:

The absence of unified standards across IoT devices, edge computing platforms, and 5G network components remains a significant barrier to widespread adoption and smooth integration of adaptive load balancing solutions. To overcome this, future work should focus on developing and promoting industry-wide standards that define common load balancing metrics, standardized APIs, and interoperable decision-making protocols. Such standardization will enable diverse systems to work together seamlessly, fostering greater collaboration, compatibility, and scalability across the 5G-IoT ecosystem.

Addressing these challenges will demand coordinated interdisciplinary efforts that bring together advances in AI, networking, distributed systems, and security. The future of scalable and resilient IoT relies on intelligent, adaptive edge infrastructure capable of meeting the complex demands of dynamic environments and diverse applications.

6. Conclusion

The rapid expansion of IoT devices, combined with the transformative power of 5G networks, demands edge computing infrastructures that are both highly efficient and scalable. This paper has emphasized the essential role of adaptive load balancing on edge gateways as a key enabler for IoT scalability in such dynamic settings. We proposed a comprehensive adaptive load balancing model that leverages real-time monitoring, AI/ML-driven predictive analytics, and intelligent decision-making to optimize resource use, reduce latency, and boost throughput across a variety of IoT applications.

Our architecture demonstrates how components like the Resource Monitor, AI/ML Prediction Engine, and Dynamic Decision Maker work together to enable proactive, context-aware load distribution among edge gateways. The model's integration with advanced 5G features such as Multi-access Edge Computing (MEC) and network slicing further enhances its performance potential. Despite these advantages, significant challenges remain—including managing dynamic environments, ensuring AI model robustness and explainability, handling resource heterogeneity, and addressing critical security and privacy issues. Future research must focus on areas such as continual learning, federated resource orchestration, lightweight explainable AI, and blockchain-based trust frameworks. By tackling these challenges, we can build an intelligently adaptive load balancing system that forms the backbone of resilient, responsive smart city, industrial, and other vital IoT ecosystems of the future.

References

[1] A. Amrullah, M. U. H. Al Rasyid, and I. Winarno, "Implementation and Analysis of IoT Communication Protocols for Crowdsensing and Crowdsourcing in Health Application," in *2021 International Electronics Symposium (IES)*, 2021, pp. 209–214. doi: 10.1109/IES53407.2021.9593999.

[2] M. E. E. Alahi *et al.*, "Integration of IoT-Enabled Technologies and Artificial Intelligence (AI) for Smart City Scenario: Recent Advancements and Future Trends," *Sensors*, vol. 23, no. 11, 2023, doi: 10.3390/s23115206.

[3] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G Industrial IoT: A Survey," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1134–1163, 2022, doi: 10.1109/OJCOMS.2022.3189013.

[4] Y. Liu, M. Peng, G. Shou, Y. Chen, and S. Chen, "Toward Edge Intelligence: Multiaccess Edge Computing for 5G and Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6722–6747, 2020, doi: 10.1109/JIOT.2020.3004500.

[5] L. Kong *et al.*, "Edge-computing-driven Internet of Things: A Survey," *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022, doi:

10.1145/3555308.

[6] S. Hamdan, M. Ayyash, and S. Almajali, “Edge-Computing Architectures for Internet of Things Applications: A Survey,” *Sensors*, vol. 20, no. 22, 2020, doi: 10.3390/s20226441.

[7] I. Rafiq, A. Mahmood, S. Razzaq, S. H. M. Jafri, and I. Aziz, “IoT applications and challenges in smart cities and services,” *J. Eng.*, vol. 2023, no. 4, p. e12262, 2023, doi: https://doi.org/10.1049/tje2.12262.

[8] S. S. Sefati and S. Halunga, “Ultra-reliability and low-latency communications on the internet of things based on 5G network: Literature review, classification, and future research view,” *Trans. Emerg. Telecommun. Technol.*, vol. 34, no. 6, p. e4770, 2023, doi: https://doi.org/10.1002/ett.4770.

[9] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, “Massive Access for 5G and Beyond,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, 2021, doi: 10.1109/JSAC.2020.3019724.

[10] Y.-J. Wu, W.-S. Hwang, C.-Y. Shen, and Y.-Y. Chen, “Network Slicing for mMTC and URLLC Using Software-Defined Networking with P4 Switches,” *Electronics*, vol. 11, no. 14, 2022, doi: 10.3390/electronics11142111.

[11] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, “Load balancing techniques in cloud computing environment: A review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 3910–3933, 2022, doi: https://doi.org/10.1016/j.jksuci.2021.02.007.

[12] M. Tang and V. W. S. Wong, “Deep Reinforcement Learning for Task Offloading in Mobile Edge Computing Systems,” *IEEE Trans. Mob. Comput.*, vol. 21, no. 6, pp. 1985–1997, 2022, doi: 10.1109/TMC.2020.3036871.

[13] P. Rebari and B. R. Killi, “Deep Learning Based Traffic Prediction for Resource Allocation in Multi-Tenant Virtualized 5G Networks,” in *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, 2023, pp. 97–102. doi: 10.1109/TENCON58879.2023.10322446.

[14] G. Bao and P. Guo, “Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges,” *J. Cloud Comput.*, vol. 11, no. 1, p. 94, 2022, doi: 10.1186/s13677-022-00377-4.

[15] Á. Brandón, M. S. Pérez, J. Montes, and A. Sanchez, “FMonE: A Flexible Monitoring Solution at the Edge,” *Wirel. Commun. Mob. Comput.*, vol. 2018, no. 1, p. 2068278, 2018, doi: https://doi.org/10.1155/2018/2068278.

[16] Y. Zhang, C. Jiang, B. Yue, J. Wan, and M. Guizani, “Information fusion for edge intelligence: A survey,” *Inf. Fusion*, vol. 81, pp. 171–186, 2022, doi: https://doi.org/10.1016/j.inffus.2021.11.018.

[17] B.-S. Roh, M.-H. Han, J.-H. Ham, and K.-I. Kim, “Q-LBR: Q-Learning Based Load Balancing Routing for UAV-Assisted VANET,” *Sensors*, vol. 20, no. 19, 2020, doi: 10.3390/s20195685.

[18] R. Casado-Vara, A. del Rey, D. Pérez-Palau, L. de-la-Fuente-Valentín, and J. M. Corchado, “Web Traffic Time Series Forecasting Using LSTM Neural Networks with Distributed Asynchronous Training,” *Mathematics*, vol. 9, no. 4, 2021, doi: 10.3390/math9040421.

[19] S. Ali *et al.*, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Inf. Fusion*, vol. 99, p. 101805, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101805.