AI-Driven Beamforming and Load Balancing in 5G Edge Networks for Smart Municipalities

Maria Trigka, Elias Dritsas, and Phivos Mylonas
University of West Attica, Egaleo 12243, Greece
{mtrigka,idritsas,mylonasf}@uniwa.gr

***Abstract*—As 5G networks evolve to support the dynamic needs of smart cities, intelligent control mechanisms at the edge are essential for ensuring adaptive connectivity and efficient resource utilization. This survey explores the intersection of artificial intelligence (AI)-driven adaptive beamforming and load balancing within edge-enabled 5G architectures with direct relevance to smart municipal infrastructures. It systematically examines the algorithmic models, input–output representations, and learning paradigms employed for these two critical functions, which are decoupled from deployment specifics. This study further analyzes how such AI modules are integrated into standardized architectures and how they map onto the control and data planes of modern radio access network (RAN) systems. Real-world applications and system-level deployments are reviewed to highlight the practical viability of these approaches in urban settings. Finally, it discusses emerging trends and outlines a forward-looking path toward scalable, AI-native infrastructure for smart municipalities.**

***Index Terms*—5G Networks, Beamforming, Load Balancing, Smart Cities, Edge AI**

I. Introduction

The proliferation of smart municipal infrastructure has introduced stringent demands for urban connectivity, edge responsiveness, and service adaptivity. Modern cities increasingly rely on dense 5G deployments to sustain heterogeneous, latency-sensitive, and bandwidth-intensive applications, from autonomous transport systems to environmental monitoring and real-time public safety services. In this context, adaptive radio access and edge coordination have emerged as key enablers of scalable connectivity in smart municipalities [1].

As 5G networks evolve into disaggregated, software-defined, and AI-enhanced architectures, the role of intelligent control mechanisms becomes pivotal. Particularly at the edge, where computational and radio resources are both constrained and dynamic, the ability to steer transmission and balance the load in real time is essential for sustaining the quality of service. Beamforming [2] and load balancing, two traditionally static or rule-based functions, must be reimagined as learning-driven, context-aware processes tailored to the spatiotemporal demands of urban environments [3].

A. Motivation and Contribution

This survey addresses the growing need for learning-enabled control in edge-native 5G systems, with particular emphasis on functions that are critical to smart city performance. By focusing on adaptive beamforming and load balancing, we explored how AI can drive these mechanisms in dynamic and resource-constrained environments. Our perspective bridges algorithmic methods with their placement and operation within real-world edge architectures, aiming to make AI integration practical and system-aware.

We present a structured taxonomy of AI models relevant to beamforming and load balancing, spanning supervised, unsupervised, and reinforcement learning (RL) approaches. Additionally, we detail how these models interface with standardized architectural frameworks, such as multi-access edge computing (MEC) and open RAN (O-RAN), and clarify their roles in the control planes and data planes. Finally, we examine practical system flows and integration pathways that support the realization of these AI-enabled functions in the context of smart municipal deployments.

B. Methodology

This survey adopts a structured and reproducible methodology to identify and analyze recent literature at the intersection of AI, beamforming, load balancing, and 5G edge architectures. Relevant peer-reviewed studies published since 2020 were retrieved from major scientific databases and digital libraries, including IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar. The inclusion focused on AI-based control in 5G systems for smart cities, particularly involving beam management, user association, and edge-native optimization. Studies limited to physical-layer design or lacking architectural relevance were also excluded. After deduplication and screening, 48 publications were selected for full-text analysis. The survey process consisted of three main steps. Initial retrieval was performed using targeted queries such as ”AI for beamforming,” ”load balancing in 5G,” ”edge intelligence,” ”smart municipalities,” and ”smart city networks.” Relevant studies were reviewed to extract information on AI models, learning objectives, and system integration aspects. Finally, the selected literature was comparatively analyzed with respect to the architectural scope, control strategies, and applicability to smart municipal systems.

The remainder of this paper is organized as follows: Section II presents the fundamental concepts of beamforming, load balancing, and edge computing in the 5G architecture. Section III analyzes AI-driven methods for adaptive beamforming and load balancing at the network edge. Section IV describes the architectural integration of AI modules into 5G edge systems. Section V highlights the real-world applications

979-8-3315-0448-9/24/31.00 ©2024 IEEE

and use cases of smart municipal systems. Section VI situates our work within the existing survey literature, discusses its broader implications, and outlines future research directions. Section VII concludes the paper.

II. Fundamental Concepts in Beamforming, Load Balancing, and Edge Architecture

Adaptive control in 5G builds on the core principles of beamforming, load balancing, and edge computing. This section outlines these foundations, with an emphasis on dense municipal deployments, in which latency, coordination, and connectivity are crucial.

A. Beamforming in 5G Architectures

Beamforming steers the transmission energy by adjusting the phase and amplitude of the antenna elements. Architectures are categorized as analog (single radio frequency (RF) chain with phase shifters), digital (per-element RF chains with full flexibility), or hybrid (reduced RF chains with analog precoding), balancing the complexity, cost, and performance in massive multiple-input–multiple-output (MIMO) setups [4] [5].

To enable spatial multiplexing, beamforming relies on predefined beam codebooks that are optimized for angular coverage and user separation. These may be discrete Fourier transform (DFT)-based, hierarchical, or adaptive to propagation conditions. Effective beamforming requires accurate channel state information (CSI). In time-division duplex (TDD) systems, reciprocity permits the reuse of uplink CSI. In frequency-division duplex (FDD) systems, the downlink CSI must be estimated and fed back, which increases the overhead. Estimation methods include pilot-based approaches, compressive sensing, and quantized feedback using codebooks [6].

The algorithm design depends on realistic antenna system and channel models. The arrays can be a uniform linear array (ULA), uniform planar array (UPA), or more complex [7] with varying beam widths and angular resolutions. Channel modeling standardized by ETSI, grounded in the 3GPP TR 38.901 framework [8], and supported by tools such as NYUSIM and QuaDRiGa, offers a detailed representation of wireless propagation from sub-6GHz to mmWave bands. By capturing path loss, fading, angular spread, and blockage across representative deployment scenarios (e.g., Urban Macro/Micro), these models enable the rigorous evaluation of physical-layer and AI-driven strategies tailored to smart municipality applications in 5G and future 6G systems [9].

B. Load Balancing in Mobile Networks

Load balancing redistributes traffic and user associations across cells to optimize resource utilization and service quality. The core indicators include resource usage, buffer occupancy, and scheduling backlog. The association decisions consider the load, signal quality, and handover constraints. A common mechanism is cell range expansion, which biases small cells to offload the macrocells. Strategies may be network-centric (centralized) or user equipment (UE)-centric (local and rule-based) [10] [11].

Optimization techniques range from convex formulations (e.g., utility maximization) to heuristics (greedy, rule-based) and combinatorial methods for discrete, constraint-aware assignment problems. In 5G systems with edge capabilities, load balancing extends to computing and data storage. MEC-aware models jointly assess RAN and edge server loads to coordinate radio-compute assignments via cross-domain signaling and a shared state [12].

C. Edge Computing in 5G Systems

Edge computing places the computation near the RAN to support low-latency services. Disaggregated RAN includes radio units (RU), distributed units (DU), and central units (CU), which separate the analog, baseband, and protocol layers. This split, per 3GPP and O-RAN, enables flexible function placement based on latency needs in real-time municipal services [13].

The MEC defines a framework for edge-hosted applications with application programming interfaces (APIs) that expose the network context (e.g., load and location). A typical MEC stack includes virtualization, orchestration, and service registries that are co-located with the DU. Latency arises from processing, fronthaul, and scheduling issues. The models distinguish between uplink/downlink paths and queuing effects. Data locality, processing at the source versus elsewhere, directly impacts the end-to-end delay and backhaul load [14] [15].

Coordination between the RAN and compute layers involves control loops and joint scheduling. Interfaces such as F1 (CU–DU) and orchestration systems (e.g., ETSI MEC, open network automation platform) enable dynamic resource allocation based on radio states and traffic forecasts [16]. A summary of the concepts is provided in Table I, which offers a comparative overview of their structure, classification, and operational roles.

III. AI for Adaptive Beamforming and Load Balancing at the Edge

AI is key to managing radio and computing resources in edge-enabled 5G networks, particularly in dynamic and service-rich municipal settings. This section reviews the learning paradigms for adaptive beamforming and load balancing, including the inputs, output, and evaluation metrics.

A. AI for Adaptive Beamforming

AI-driven beamforming aims to infer optimal transmission parameters based on observed radio channel conditions. These approaches are broadly categorized as supervised, reinforcement, and unsupervised learning [17].

In supervised learning, models are trained to map channel features to beamforming actions using labeled data. Typical tasks include beam index prediction, beam direction classification, and precoder regression. Convolutional neural networks (CNNs), recurrent networks, and transformer architectures

TABLE I
KEY CONSTRUCTS IN BEAMFORMING, LOAD BALANCING, AND EDGE COMPUTING IN 5G NETWORKS.

Concept	Key Elements	Categories	Operational Role
Beamforming	Antenna arrays, signal phase/amplitude control, CSI	Analog / Digital / Hybrid	Beam steering, spatial multiplexing, interference control
CSI Acquisition	Pilot signals, feedback schemes	TDD/FDD, compressive sensing, codebook-based	Beam selection, link adaptation
Beam Codebooks	Predefined beam sets, angular design	DFT, hierarchical, data-driven	Efficient beam search, quantization
Load Balancing	Cell load metrics, user association rules	Network-centric / UE-centric	Traffic redistribution, congestion mitigation
Optimization Methods	Resource allocation logic	Convex, heuristic, combinatorial	User-cell mapping, load fairness
MEC-Aware Balancing	Joint radio-compute status	Single / dual-tier coordination	Offloading, edge-aware user assignment
Edge Architecture	RU/DU/CU split, MEC stack	3GPP, O-RAN-based	Distributed processing, compute orchestration
Latency Factors	Fronthaul, scheduling, compute delays	Uplink/downlink split, queueing models	End-to-end delay, service-level agreement compliance

have been employed to process spatial channel representations such as CSI matrices and channel power maps [18].

RL methods frame beam adaptation as a sequential decision-making process in which an agent selects beam actions based on environmental feedback. Model-free algorithms, such as the deep deterministic policy gradient (DDPG) and proximal policy optimization (PPO), are used to optimize beam trajectories with respect to link quality and spectral efficiency. RL is particularly suited to environments with dynamic user movement and partial observability [19].

Unsupervised techniques have been applied for beam clustering, codebook compression, and feature extraction. Methods such as k-means, autoencoders, and manifold learning identify low-dimensional structures in the CSI data to support fast beam selection or reduce feedback overhead [20].

The input features typically include raw or preprocessed CSI vectors, spatial parameters such as the angle of arrival (AoA) or angle of departure (AoD), and the full spatial channel matrix. The output targets may be discrete beam indices, continuous precoding vectors, or codebook selection masks, depending on the system constraints and learning objectives. The model performance was assessed in terms of the convergence rate, inference latency, computational complexity, and generalization across channel conditions, antenna configurations, and user mobility patterns [21] [22].

B. AI for Load Balancing at the Edge

AI-based load balancing addresses the dynamic assignment of users to network and computing resources under time-varying traffic and topology conditions. Machine learning (ML) models are used for user association by mapping individual or group-level features to the serving node assignments. Graph neural networks (GNNs) are effective in capturing spatial dependencies and network topology, thereby enabling relational reasoning across cells and user distributions [23].

RL techniques, including Q-learning and actor-critic variants, have been employed for traffic steering, edge offloading, and resource allocation. Agents learn the association or migration policies that adapt to load gradients and interference patterns. The RL framework supports both discrete action spaces (e.g., cell selection) and continuous decision-making (e.g., bandwidth allocation) [24].

Multi-agent RL (MARL) addresses the coordination among distributed agents representing access points and edge nodes. Policies are optimized either independently or jointly using centralized training paradigms [25]. Traffic forecasting was incorporated as an auxiliary learning task to inform proactive load balancing. Time-series models and temporal GNNs predict short-term demand evolution, improving congestion anticipation and enabling preemptive resource reallocation. Learning methods are evaluated based on scalability, stability under dynamic conditions, convergence behavior, and ability to maintain balanced load distributions under heterogeneous demand profiles [26]. Table II summarizes the main learning paradigms applied to adaptive beamforming and load balancing, along with their corresponding input representations and model outputs. The taxonomy distinguishes between supervised, reinforcement, and unsupervised approaches, highlighting their relevance to decision-making tasks in dynamic 5G environments.

IV. INTEGRATION ARCHITECTURES

Deploying AI for beamforming and load balancing in 5G requires clear architectural support across the radio and compute layers. In edge-native smart cities, accurate module placement, interfaces, and data flow are critical. This section describes the functional split, integration layers, and learning pipelines that enable responsive AI control.

A. Functional Decomposition

Architecturally, 5G networks separate the control and data planes. The control plane governs signaling, policy enforcement, and orchestration, whereas the data plane handles user traffic forwarding and physical-layer execution. This separation allows the deployment of AI controllers as logically distinct components within the network hierarchy [27].

The placement of AI control modules may follow centralized, edge-native, or hybrid schemes. In the centralized case, model inference and policy decisions are executed in the cloud or core domains with global network visibility. Edge-native placement brings decision logic closer to the data source, enabling a lower-latency reaction. Hybrid schemes combine cloud-scale learning with local execution by partitioning model responsibilities across network tiers [28].

B. Standardized Architectures

The O-RAN architecture provides a modular framework for integrating AI functionality into 5G RANs through well-defined interfaces. Central to this architecture is the near-

TABLE II
AI LEARNING PARADIGMS FOR BEAMFORMING AND LOAD BALANCING IN 5G NETWORKS.

Task	Learning Paradigms	Input Features	Model Outputs
Adaptive Beamforming	Supervised (e.g., CNN, transformer); RL (e.g., DDPG, PPO); Unsupervised (e.g., clustering, autoencoders)	CSI matrices, spatial channel maps, AoA/AoD statistics	Beam indices, precoders, codebook entries, latent beam clusters
Load Balancing	Supervised (e.g., GNN); RL (Q-learning, actor-critic); MARL; Auxiliary forecasting models	Cell load, user location, traffic state, topology graphs, temporal demand traces	Association policies, offloading actions, coordination strategies, load predictions

real-time RAN intelligent controller (near-RT RIC), which hosts xApps and modular applications responsible for local inference and control within strict latency budgets. The non-real-time RIC, which is located at higher layers, supports offline training, analytics, and policy optimization [29].

MEC integration enables the co-location of computing resources with baseband units for latency-sensitive execution. MEC nodes expose the radio context to applications through service APIs and act as hosts for deploying inference runtimes that interact with RAN components [30]. Softwarized RAN paradigms, such as software-defined RAN (SD-RAN) and network function virtualization (NFV), further decouple hardware from logic, allowing programmable control loops to embed AI modules using virtualized network functions (VNF). These frameworks enable orchestration platforms to dynamically allocate resources, load models, and coordinate inference workflows across distributed infrastructures [31].

C. Data Flow and Training Pipelines

The integration of AI functionality requires the definition of data pipelines that connect measurement sources, inference models, and actuation points. Typical flow sequences consist of raw radio features (e.g., CSI), buffer states, and scheduling metrics collected at the RU or DU, preprocessed, and forwarded to model execution engines. The output actions are then applied via configuration interfaces to beamformers, schedulers, or user association controllers [32].

Training workflows can operate in online or offline modes. In the online mode, inference models are updated in situ with streaming data under resource and latency constraints. Offline training pipelines collect data over time, process them in centralized environments, and deploy updated model instances to network edge nodes using orchestration agents [33].

The interface specifications govern the communication between the model modules and network functions. These include standardized APIs (e.g., O-RAN, A1 (policy management), and E2 (real-time control) interfaces), telemetry buses, and feedback loops that enable closed-loop control. Proper synchronization of data, models, and decision outputs is essential for maintaining consistency across layers and ensuring deterministic behavior under dynamic network conditions [34]. Table III outlines the key architectural elements involved in the AI integration workflows, including their functional responsibilities and corresponding interfaces within the 5G infrastructure.

V. APPLICATIONS AND USE CASES

In dense city zones, where angular dispersion and user mobility are prominent owing to vehicular traffic and crowd movement, edge-deployed beam selection models have been shown to enhance link robustness and continuity. Smart transportation corridors, surveillance grids, and responsive infrastructures benefit from real-time CSI processing at DUs, where inference engines select optimal beam indices based on spatial radio signatures. Implementations using hierarchical codebooks and AoA estimators were validated in OpenAirInterface and srsRAN environments, with near-RT xApps deployed on RIC platforms. These configurations reduce the beam-switching delay and improve resilience to blockage, particularly in dynamic municipal deployments, such as public squares and transit hubs [35].

In heterogeneous network topologies typical of urban environments, where macro base stations support wide-area coverage while small cells serve hotspots (e.g., stations, markets, and municipal buildings), AI-driven user association dynamically redistributes the loads. Graph-based neural models leverage live snapshots of the network topology and buffer telemetry to optimize user-cell mappings. MARL methods trained in 3GPP-compliant simulators, such as Simu5G or ns-3 mmwave, have demonstrated decentralized adaptation to traffic shifts, enabling seamless mobility support and interference-aware handover management. These approaches directly address the fluctuating usage patterns of smart municipal services, from emergency dispatch systems to event-driven public connectivity [36] [37].

Advanced use cases in city-level deployments involve the joint optimization of beamforming and task offloading across the radio and edge compute layers. For instance, in real-time analytics systems supporting urban surveillance or traffic regulation, AI agents interface with both RAN controllers and compute orchestrators to align the beam direction with the MEC server assignment. Such joint decision frameworks have been realized via ETSI MEC-compliant stacks over containerized platforms (e.g., Kubernetes with Akraino profiles). Predictive policies account for compute saturation and bursty traffic, maintaining throughput and latency targets under mixed workload conditions that are common in smart city control loops [38] [39].

Large-scale testbeds, such as POWDER-RENEW and COSMOS, have validated the feasibility of deploying such AI-enabled mechanisms in settings simulating urban density and live mobility. These platforms support programmable SD-RANs and expose real-time APIs for model inference and data feedback, closely reflecting the operational environments of the smart municipalities. Metrics, including average link rate, association persistence, and inference latency, have been used for evaluation, although unified benchmarking across

TABLE III
ARCHITECTURAL LAYERS AND INTERFACES FOR AI INTEGRATION IN 5G SYSTEMS.

Layer / Module	**Role**	**Relevance to AI Integration**
Control vs Data Plane	Logical separation of signaling and forwarding	Control plane hosts decision loops; data plane executes AI-configured actions
O-RAN RIC (Near/Non-RT)	Real-time control (xApps) and policy management	xApps run inference; A1/E2 interfaces manage training and orchestration
MEC / NFV Stack	Edge compute and virtualization layer	Hosts model runtimes, allocates resources dynamically
Data Pipeline	Collection and preprocessing of network features	Supplies, CSI, traffic/load data to inference engines
Model Runtime / Server	Model inference and lifecycle handling	Executes, updates, and coordinates AI modules at edge or core
Standard Interfaces	Inter-component connectivity (A1, E2, F1)	Support telemetry, policy transfer, and closed-loop control

deployments remains an open issue. Public datasets, such as DeepMIMO, NYU Wireless, and O-RAN logs, have served as training and validation baselines for emulating city-scale traffic and channel dynamics [40] [41].

VI. Discussion and Future Trends

Recent surveys have examined the role of mobile networks in smart cities, emphasizing architecture, sustainability, and service integration. One such study outlines 5G/6G paradigms but does not address beamforming or resource control mechanisms [42], while another remains high-level and omits AI-driven subsystem modeling [43]. In contrast, our survey focuses on system-level interfaces and learning models for adaptive beamforming and load balancing at the edge, offering a more technical and deployment-oriented perspective.

Other reviews explore Beyond 5G slicing architectures [44], but concentrate on logical isolation rather than physical-layer control. Work on AI-based handovers and load optimization covers ML policies in dense deployments [45], yet lacks an architectural integration flow. Our contribution addresses this gap by mapping AI modules onto standardized frameworks (e.g., O-RAN and MEC) to enable coordinated radio-compute resource management.

Beamforming-focused surveys present detailed learning taxonomies [46], but typically abstract from deployment and latency constraints. Similarly, RL-based reconfigurable intelligent surface control offers algorithmic insights [47], although its practical relevance is limited. Our work remains grounded in near-term deployable mechanisms within 5G edge systems.

In distributed ML for wireless, prior work emphasizes federated inference and model placement [48], but treats beamforming and load balancing marginally. We position these functions as central to the AI–edge co-design loop.

For practitioners, our system-level framing offers a clear mapping between AI control logic and edge architectures, thereby supporting scalable and adaptive operations across urban networks. The survey highlights open problems in decentralized control, generalization, and training under tight latency constraints.

Looking ahead, future trends are expected to center around federated RL for multi-cell coordination, the integration of semantic communication layers for context-aware transmission, and embedding AI functions within hardware-constrained platforms. Moreover, as smart city environments become increasingly data-intensive and multimodal, hybrid control policies that combine symbolic reasoning with neural methods may emerge as promising avenues. Finally, sustainability concerns will drive the design of energy-efficient AI agents capable of self-pruning, compression, and real-time edge retraining.

VII. Conclusions

This survey examines how AI-driven control functions, specifically adaptive beamforming and load balancing, can be systematically integrated into edge-native 5G infrastructures that underpin smart municipal services. By disentangling algorithmic mechanisms from deployment architectures, we provide a unified view of how learning-based models interact with disaggregated RAN stacks, MEC platforms, and standardized interfaces such as O-RAN.

Our contribution lies in bridging ML model taxonomies with practical integration flows, offering a deployment-oriented perspective that informs both system designers and urban connectivity planners. This synthesis is particularly relevant for smart municipalities seeking scalable, responsive, and context-aware network controls. Future research is expected to emphasize federated edge learning, real-time inference under resource constraints, and co-optimization of radio and compute functions as key enablers of resilient and adaptive 5G systems.

References

[1] C. Yang, P. Liang, L. Fu, G. Cui, F. Huang, F. Teng, and Y. A. Bangash, "Using 5g in smart cities: A systematic mapping study," *Intelligent Systems with Applications*, vol. 14, p. 200065, 2022.

[2] S. Enahoro, S. C. Ekpo, M. Uko, F. Elias, and S. Alabi, "Integrating iot with adaptive beamforming for enhanced urban sensing in smart cities," *IEEE Access*, 2025.

[3] P. Tarafder and W. Choi, "Deep reinforcement learning-based coordinated beamforming for mmwave massive mimo vehicular networks," *Sensors*, vol. 23, no. 5, p. 2772, 2023.

[4] J. Zhang, X. Yu, and K. B. Letaief, "Hybrid beamforming for 5g and beyond millimeter-wave systems: A holistic view," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 77–91, 2020.

[5] M. Majidzadeh, J. Kaleva, N. Tervo, H. Pennanen, A. Tölli, and M. Latva-aho, "Hybrid beamforming for mm-wave massive mimo systems with partially connected rf architecture," *Wireless Personal Communications*, vol. 136, no. 4, pp. 1947–1979, 2024.

[6] Q. Ziao and Y. Haifan, "A review of codebooks for csi feedback in 5g new radio and beyond," *China Communications*, vol. 22, no. 2, pp. 112–127, 2025.

[7] S. Khan, T. Mazhar, T. Shahzad, A. Bibi, W. Ahmad, M. A. Khan, M. M. Saeed, and H. Hamam, "Antenna systems for iot applications: a review," *Discover Sustainability*, vol. 5, no. 1, p. 412, 2024.

[8] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 18)," ETSI, Tech. Rep. TR 138.901 V18.0.0, 2024. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/18.00.00_60/tr_138901v180000p.pdf

[9] L. Pang, J. Zhang, Y. Zhang, X. Huang, Y. Chen, and J. Li, "Investigation and comparison of 5g channel models: From quadriga, nyusim, and mg5g perspectives," *Chinese Journal of Electronics*, vol. 31, no. 1, pp. 1–17, 2022.

[10] M. K. Hasan, T. C. Chuah, A. A. El-Saleh, M. Shafiq, S. A. Shaikh, S. Islam, and M. Krichen, "Constriction factor particle swarm optimization based load balancing and cell association for 5g heterogeneous networks," *Computer Communications*, vol. 180, pp. 328–337, 2021.

[11] M. J. Alam, R. Chugh, S. Azad, and M. R. Hossain, "Optimizing cell association in 5g and beyond networks: a modified load-aware biased technique," *Telecommunication Systems*, vol. 87, no. 3, pp. 731–742, 2024.

[12] W. Chen, Y. Zhu, J. Liu, and Y. Chen, "Enhancing mobile edge computing with efficient load balancing using load estimation in ultra-dense network," *Sensors*, vol. 21, no. 9, p. 3135, 2021.

[13] W. Azariah, F. A. Bimo, C.-W. Lin, R.-G. Cheng, N. Nikaein, and R. Jana, "A survey on open radio access networks: Challenges, research directions, and open source approaches," *Sensors*, vol. 24, no. 3, p. 1038, 2024.

[14] P. Ranaweera, A. D. Jurcut, and M. Liyanage, "Survey on multi-access edge computing security and privacy," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1078–1124, 2021.

[15] G. Nencioni, R. G. Garroppo, and R. F. Olimid, "5g multi-access edge computing: A survey on security, dependability, and performance," *IEEE Access*, vol. 11, pp. 63 496–63 533, 2023.

[16] R. Cannata, H. Sun, D. M. Dumitriu, and H. Hassanieh, "Towards seamless 5g open-ran integration with webassembly," in *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, 2024, pp. 121–131.

[17] C. M. Andras, G. Barb, and M. Otesteanu, "Comparative analysis of beamforming techniques and beam management in 5g communication systems," *Sensors*, vol. 25, no. 15, p. 4619, 2025.

[18] E. Chatzoglou and S. K. Goudos, "Beam-selection for 5g/b5g networks using machine learning: A comparative study," *Sensors*, vol. 23, no. 6, p. 2967, 2023.

[19] A. Y. Sarhan, O. A. Abdullah, H. Al-Hraishawi, and F. S. Alsubaei, "Reinforcement learning-driven secrecy energy efficiency maximization in ris-enabled communication systems," *IEEE Access*, 2025.

[20] M. Baur, M. Würth, M. Koller, V.-C. Andrei, and W. Utschick, "Csi clustering with variational autoencoding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5278–5282.

[21] S. Lavdas, P. K. Gkonis, E. Tsaknaki, L. Sarakis, P. Trakadas, and K. Papadopoulos, "A deep learning framework for adaptive beamforming in massive mimo millimeter wave 5g multicellular networks," *Electronics*, vol. 12, no. 17, p. 3555, 2023.

[22] M. Alrabeiah, Y. Zhang, and A. Alkhateeb, "Neural networks based beam codebooks: Learning mmwave massive mimo beams that adapt to deployment and hardware," *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 3818–3833, 2022.

[23] O. Orhan, V. N. Swamy, T. Tetzlaff, M. Nassar, H. Nikopour, and S. Talwar, "Connection management xapp for o-ran ric: A graph neural network and reinforcement learning approach," in *2021 20th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2021, pp. 936–941.

[24] Z. Wu, Z. Jia, X. Pang, and S. Zhao, "Deep reinforcement learning-based task offloading and load balancing for vehicular edge computing," *Electronics*, vol. 13, no. 8, p. 1511, 2024.

[25] A. Alizadeh, B. Lim, and M. Vu, "Multi-agent q-learning for real-time load balancing user association and handover in mobile networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 9001–9015, 2024.

[26] S. Liu, M. He, Z. Wu, P. Lu, and W. Gu, "Spatial–temporal graph neural network traffic prediction based load balancing with reinforcement learning in cellular networks," *Information Fusion*, vol. 103, p. 102079, 2024.

[27] M. Al Shinwan, L. Abualigah, T.-D. Huy, A. Younes Shdefat, M. Altalhi, C. Kim, S. El-Sappagh, M. Abd Elaziz, and K. S. Kwak, "An efficient 5g data plan approach based on partially distributed mobility architecture," *Sensors*, vol. 22, no. 1, p. 349, 2022.

[28] S. Tuli, F. Mirhakimi, S. Pallewatta, S. Zawad, G. Casale, B. Javadi, F. Yan, R. Buyya, and N. R. Jennings, "Ai augmented edge and fog computing: Trends and challenges," *Journal of Network and Computer Applications*, vol. 216, p. 103648, 2023.

[29] M. El-Hajj, "Secure and trustworthy open radio access network (o-ran) optimization: A zero-trust and federated learning framework for 6g networks," *Future Internet*, vol. 17, no. 6, p. 233, 2025.

[30] R. Xavier, R. S. Silva, M. Ribeiro, W. Moreira, L. Freitas, and A. Oliveira-Jr, "Integrating multi-access edge computing (mec) into open 5g core," in *Telecom*, vol. 5, no. 2. MDPI, 2024, pp. 433–450.

[31] F. Chiti, S. Morosi, and C. Bartoli, "An integrated software-defined networking–network function virtualization architecture for 5g ran–multi-access edge computing slice management in the internet of industrial things," *Computers*, vol. 13, no. 9, p. 226, 2024.

[32] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Pèrez, and C. Verikoukis, "Explainable ai in 6g o-ran: A tutorial and survey on architecture, use cases, challenges, and future research," *IEEE Communications Surveys & Tutorials*, 2024.

[33] T. Zeng, X. Zhang, J. Duan, C. Yu, C. Wu, and X. Chen, "An offline-transfer-online framework for cloud-edge collaborative distributed reinforcement learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 35, no. 5, pp. 720–731, 2024.

[34] M. Polese, L. Bonati, S. D'oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376–1411, 2023.

[35] M. Silva, J. P. Fonseca, D. P. Abreu, P. Martins, P. Duarte, R. Barbosa, B. Mendes, J. Silva, A. Goes, M. Araujo *et al.*, "O-ran and ric compliant solutions for next generation networks," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2023, pp. 1–7.

[36] I. Alablani and M. J. Alenazi, "Dqn-gnn-based user association approach for wireless networks," *Mathematics*, vol. 11, no. 20, p. 4286, 2023.

[37] P. Ramesh, P. Bhuvaneswari, V. Dhanushree, G. Gokul, and S. Sahana, "User association-based load balancing using reinforcement learning in 5g heterogeneous networks," *The Journal of Supercomputing*, vol. 81, no. 1, p. 328, 2025.

[38] H. Zhang, Z. Tian, L. Zeng, L. Lu, S. Qiao, S. Chen, and X. Liu, "Distributed multi-agent reinforcement learning approach for multi-server multi-user task offloading," *IEEE Internet of Things Journal*, 2025.

[39] P. Liu, Z. Fei, X. Wang, Y. Zhou, Y. Zhang, and F. Liu, "Joint beamforming and offloading design for integrated sensing, communication, and computation system," *IEEE Transactions on Vehicular Technology*, 2025.

[40] Z. Kostic, A. Angus, Z. Yang, Z. Duan, I. Seskar, G. Zussman, and D. Raychaudhuri, "Smart city intersections: Intelligence nodes for future metropolises," *Computer*, vol. 55, no. 12, pp. 74–85, 2022.

[41] J. Breen, A. Buffmire, J. Duerig, K. Dutt, E. Eide, M. Hibler, D. Johnson, S. K. Kasera, E. Lewis, D. Maas *et al.*, "Powder: Platform for open wireless data-driven experimental research," in *Proceedings of the 14th International Workshop on Wireless Network Testbeds, Experimental evaluation & Characterization*, 2020, pp. 17–24.

[42] S. Islam, A. Z. Abdulsalam, B. A. Kumar, M. K. Hasan, R. Kolandaisamy, and N. Safie, "Mobile networks toward 5g/6g: Network architecture, opportunities and challenges in smart city," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 3082–3093, 2024.

[43] M. J. Shehab, I. Kassem, A. A. Kutty, M. Kucukvar, N. Onat, and T. Khattab, "5g networks towards smart and sustainable cities: A review of recent developments, applications and future perspectives," *IEEe Access*, vol. 10, pp. 2987–3006, 2021.

[44] W. Rafique, J. R. Barai, A. O. Fapojuwo, and D. Krishnamurthy, "A survey on beyond 5g network slicing for smart cities applications," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 1, pp. 595–628, 2024.

[45] C. Chabira, I. Shayea, G. Nurzhaubayeva, L. Aldasheva, D. Yedilkhan, and S. Amanzholova, "AI-driven handover management and load balancing optimization in ultra-dense 5G/6G cellular networks," *Technologies*, vol. 13, no. 7, p. 276, 2025.

[46] D. d. S. Brilhante, J. C. Manjarres, R. Moreira, L. de Oliveira Veiga, J. F. de Rezende, F. Muller, A. Klautau, L. Leonel Mendes, and F. A. P. de Figueiredo, "A literature survey on AI-aided beamforming and beam management for 5G and 6G systems," *Sensors*, vol. 23, no. 9, p. 4359, 2023.

[47] A. A. Puspitasari and B. M. Lee, "A survey on reinforcement learning for reconfigurable intelligent surfaces in wireless communications," *Sensors*, vol. 23, no. 5, p. 2554, 2023.

[48] O. Nassef, W. Sun, H. Purmehdi, M. Tatipamula, and T. Mahmoodi, "A survey: Distributed machine learning for 5g and beyond," *Computer Networks*, vol. 207, p. 108820, 2022.