smart cities

MDPI

Review

Traffic Scheduling and Resource Allocation for Heterogeneous Services in 5G New Radio Networks: A Scoping Review

Ntunitangua René Pindi [1,2,3] and Fernando J. Velez [1,2,*]

[1] Instituto de Telecomunicações, Faculdade de Engenharia, Universidade da Beira Interior, Calçada Fonte do Lameiro, 6201-001 Covilhã, Portugal; rene.pindi@ubi.pt
[2] Departmento de Engenharia Electromecânica, Faculdade de Engenharia, Universidade da Beira Interior, Calçada Fonte do Lameiro, 6201-001 Covilhã, Portugal
[3] Departamento de Engenharia e Telecomunicações, Instituto Superior Politécnico de Ndalatando, Ndalatando, Angola
* Correspondence: fjv@ubi.pt

Highlights

What are the main findings?

- The study highlights the importance of research on mMTC reliability in 5G networks, highlighting the need for AI-driven methodologies to balance latency, throughput, and energy efficiency.
- Currently, research mainly focuses on eMBB-URLLC coexistence (81.25%), but mMTC integration is underexplored, highlighting gaps in addressing its scalability and reliability for future 6G applications.

What is the implication of the main finding?

- Efficient resource allocation and frame-scheduling methods are crucial for reconciling conflicting QoS demands in multi-service 5G networks.
- Future research should prioritize tri-service coexistence to support complex applications in automated factories, telemedicine, and intelligent urban infrastructures.

Abstract

The rapid evolution of 5G New Radio networks has introduced a wide range of services with diverse requirements, complicating their coexistence within the shared radio spectrum and posing challenges in traffic scheduling and resource allocation. This study aims to analyze and categorize the methods, approaches, and techniques proposed to ensure efficient joint and dynamic packet scheduling and resource allocation among heterogeneous services—namely eMBB, URLLC, and mMTC—in 5G and beyond, with a focus on Quality of Service and user satisfaction. This scoping review draws from publications indexed in IEEE Xplore and Scopus and synthesizes the most relevant evidence related to packet scheduling across heterogeneous services, highlighting key approaches, core performance metrics, and emerging trends. Following the PRISMA-ScR methodology, 48 out of an initial 140 articles were included for explicitly addressing coexistence, scheduling, and resource allocation. The findings reveal a research emphasis on eMBB and URLLC coexistence, while integration with mMTC remains underexplored. Moreover, the evidence suggests that hybrid and deep learning-based approaches are particularly promising for tackling coexistence and resource management challenges in future mobile networks.

Keywords: frame scheduling; coexistence; eMBB; URLLC; mMTC; 5G

Academic Editor: Pierluigi Siano

Received: 13 July 2025
Revised: 2 October 2025
Accepted: 4 October 2025
Published: 10 October 2025

Citation: Pindi, N.R.; Velez, F.J. Traffic Scheduling and Resource Allocation for Heterogeneous Services in 5G New Radio Networks: A Scoping Review. *Smart Cities* **2025**, *8*, 168. https://doi.org/10.3390/smartcities8050168

1. Introduction

Fifth-generation mobile networks, known as 5G New Radio (5G NR), deliver significantly faster and more secure communications than their predecessors. The International Telecommunication Union (ITU) defines three principal service categories for 5G NR: enhanced Mobile Broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine type communication (mMTC).

The eMBB service targets high-data-rate applications that demand stable wireless links and medium-to large-area coverage, such as high-definition video streaming, cloud computing, and augmented virtual reality. In contrast, URLLC imposes stringent requirements on latency, reliability, and data rate, making it critical for real-time domains such as telesurgery, industrial automation (Industry 4.0), and vehicular safety. Finally, mMTC supports extremely dense device deployments with comparatively modest performance requirements [1–4].

The heterogeneous requirements of these services have motivated the development of sophisticated radio-resource-management strategies. URLLC typically receives the highest priority, achieving data rates of approximately 20 Mb/s with End-to-End (E2E) latencies below 1 ms. eMBB ranks second, providing up to 100 Mb/s with approximately 10 ms E2E latency, whereas mMTC accommodates massive connectivity at approximately 100 kb/s while tolerating higher latencies [5].

Within 5G NR, packet scheduling orchestrates the allocation and temporal use of radio resources, thereby influencing overall efficiency, fairness, latency, and user experience. Allocation reserves the necessary resources, multiplexing aggregates multiple data flows into a shared physical medium, and scheduling determines their temporal order [6–9].

The simultaneous operation of diverse services presents a significant challenge in 5G NR due to the variances in Quality of Service (QoS) demands across different traffic categories. The academic discourse delineates five principal categories of potential solutions: methodologies centered on multiplexing, QoS allocation, Machine Learning (ML), network slicing, and centralized Radio Access Network (RAN) frameworks. Nonetheless, there persists a conspicuous deficiency in comprehensive investigations that concurrently explore the coexistence of eMBB, URLLC, and mMTC traffic [5].

Therefore, this scoping review synthesizes current evidence on joint packet scheduling and resource allocation for heterogeneous 5G and post-5G services. We evaluated key performance indicators, including the E2E latency, reliability, spectral efficiency, throughput, fairness, and energy consumption, highlighting the prevailing methods, outstanding challenges, and research opportunities.

This review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines, which are designed to map existing evidence on a specific topic and to identify key concepts, theoretical frameworks, sources, and gaps in knowledge. The PRISMA-ScR framework enhances methodological rigor and transparency by providing a standardized checklist and flow diagram [10,11]. Abstract screening was assisted by Rayyan AI, a web- and mobile-based platform that streamlines collaboration during scoping reviews and is currently used by more than 700,000 researchers worldwide [12].

This scoping review aims to determine the methods, approaches, and techniques that have been proposed to achieve efficient joint and dynamic packet scheduling and resource allocation for heterogeneous services (eMBB, URLLC, and mMTC) in 5G and post-5G networks while maintaining QoS and user satisfaction.

The new findings of this study are not the original experimental results generated by the authors, but rather conclusions and perspectives derived from a scoping analysis of the existing literature. This analysis shows that most of the focus is on service combinations

between eMBB and URLLC (approximately 81.25%), while only 18.75% address combinations involve the mMTC service. Furthermore, we identified a clear neglect of the reliability of mMTC in underrepresented studies. Although reliability is not an intrinsic metric of mMTC, it remains a crucial performance indicator, potentially impacting the continuous connectivity of critical sensors and devices in areas such as smart cities, Industry 4.0, the Internet of Things (IoT), and future Sixth Generation (6G) applications.

The remainder of this article is organized as follows: Section 2 details the review methodology; Section 3 presents the results; Section 4 discusses the findings in relation to the research questions; and Section 5 concludes the paper.

2. Materials and Methods

This scoping review was conducted in compliance with the PRISMA-ScR criteria. Figure 1 presents the flowchart of the selection process in Section 3, while the completed PRISMA-ScR checklist is provided as Supplementary Materials. Although the protocol for this review was not registered in databases such as Open Science Framework (OSF) or International Platform of Registered Systematic Review and Meta-analysis Protocols (INPLASY), all steps were thoroughly documented throughout this paper.

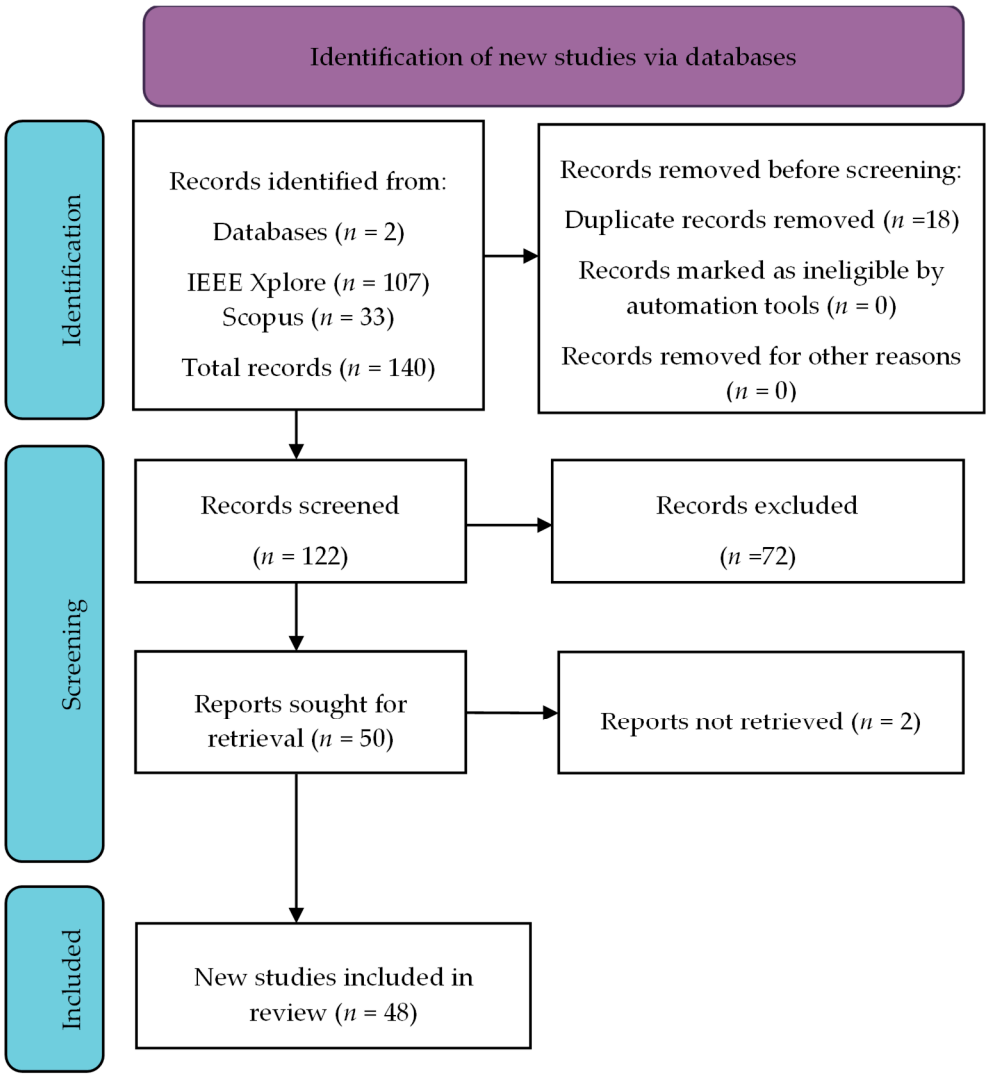

Figure 1. PRISMA flow diagram.

The review analyzes the current international literature on traffic scheduling for heterogeneous services within 5G NR networks, based on publications retrieved from the IEEE Xplore and Scopus databases between January 2019 and December 2024. Only peer-reviewed publications written in English were included. The searches were conducted using customized keyword combinations. Table 1 lists the keywords and the two search strings employed: "scheduling," "coexistence," "eMBB," "URLLC," "mMTC," and "5G" to identify relevant articles.

Table 1. Keywords and search strings used for consultation.

Keywords	Date	Cod_String	Advanced Query	Filter	Databases	Nº
Scheduling, Coexistence, eMBB, URLLC, mMTC and 5G	10 January 2025	String1	("All Metadata": Scheduling) AND ("Abstract": Coexistence) AND (("Document Title": eMBB) AND ("Document Title": URLLC)) OR ("Document Title": mMTC) AND ("Abstract": 5G)	2019–2024	IEEE Xplore	107
		String2	ALL (scheduling) AND ABS (coexistence) AND (TITLE (eMBB) AND TITLE (URLLC)) OR TITLE (mmtc) AND ABS (5g) AND PUBYEAR > 2018 AND PUBYEAR < 2025	2019–2024	Scopus	33

In the IEEE Xplore database, search string 1 has been applied to the advanced search interface with a 2019–2024 publication filter, yielding 107 records. For the Scopus database, the same procedure has been repeated using Search String 2, which has returned 33 records.

The results from both databases have been exported independently in Research Information Systems (RIS) format and then imported into Rayyan AI, a web-based platform for scoping reviews and meta-analyses, to streamline the selection, screening, and organization of studies. Rayyan has also accelerated the review process and improved consistency in inclusion and exclusion-related decisions.

A total of 140 records have been imported into Rayyan. After automatic duplicate detection, 18 records were removed, leaving 122 unique titles and abstracts for screening. Of these, 72 records have been excluded and 50 have been retained for full-text assessment based on the inclusion and exclusion criteria described in the next section.

To ensure maximum objectivity and transparency in the selection process, conducted by a single researcher (who conducted the review), the following procedures have been adopted: The Rayyan tool has been set to Blind mode, meaning that the reviewer's decisions for inclusion or exclusion were recorded without access to information such as the author's name, the journal, or the affiliation, thus reducing potential unconscious biases; The inclusion and exclusion criteria (detailed in Sections 2.1.1 and 2.1.2) have been established a priori and applied strictly in a binary manner (yes/no) to each record during screening. After a two-week interval, the researcher re-evaluated a random sample of 10% of the records (n = 12) to verify the internal consistency of their own decisions (intra-observer reliability). A 100% agreement was observed in this re-evaluation, indicating a high consistency in the application of the criteria; The complete workspace of the review in Rayyan, including the 122 records and their respective classifications and tags, has been archived. This material is available for audit and verification by any interested party, upon request to the authors.

2.1. Inclusion and Exclusion Criteria for the Scoping Review

For study selection, we targeted articles that examined the fundamental and interdependent aspects of 5G NR networks, namely the coexistence of heterogeneous services, packet scheduling, multiplexing, and resource allocation, and, by implication, optimization techniques. Achieving the performance objectives of 5G NR requires treating these elements not as isolated topics but as mutually dependent components of a unified system, thereby enabling the efficient, precise, and coordinated management of network resources.

2.1.1. Inclusion Criteria

Studies were eligible for inclusion if they:

- First bullet Explicitly examined packet-traffic scheduling that accounts for the concurrent operation of all three 5G NR service categories (eMBB, URLLC, and mMTC) or any of their pairwise combinations;
- Described concrete scheduling techniques for managing service coexistence, such as puncturing (i.e., the temporary interruption or preemption of ongoing transmissions), overlay, network slicing, hybrid schemes, or machine-learning-based approaches;
- Focused on multiplexing mechanisms that merge heterogeneous traffic onto shared radio resources;
- Proposed algorithms or strategies for flexible and efficient allocation of Resource Blocks (RBs), transmit power, or time among multiple users and services;
- Optimization of key QoS metrics for different service types, including latency, reliability, throughput, and spectral efficiency;
- Analytical models, simulations, and experimental evidence are provided to evaluate the effectiveness of the proposed scheduling method.

2.1.2. Exclusion Criteria

Studies were excluded if they:

- It does not address at least one of the following core concepts: service coexistence, traffic scheduling, multiplexing, or resource allocation in 5G or post-5G networks;
- It focuses exclusively on a single service type (eMBB, URLLC, or mMTC), without considering coexistence or joint scheduling with other services;
- They provided only theoretical discussions or conceptual models, excluding literature or scoping reviews, without accompanying simulations or experimental validation;
- Targeted higher network layers (e.g., applications and services) without describing their interactions with radio resource-scheduling mechanisms.

3. Results

After full reading of the articles and application of the predefined criteria, two additional articles were excluded due to non-compliance, resulting in 48 articles included in this scoping review. Figure 1 illustrates the flowchart and protocol used to identify and select the articles. It is worth noting that among the selected articles, two were literature reviews.

Table 2 summarizes 48 articles published between 2019 and 2024. The table comprises eight columns designed to capture the key data extracted from each study. At a minimum, all selected articles addressed the challenges of service coexistence, packet scheduling, and resource allocation in 5G and post-5G networks while considering the distinct requirements of each service category.

Table 2. Summary of included articles.

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[13]	eMBB and mMTC	Coexistence and power allocation	Game theory (specifically Stackelberg-Nash Game combined with Mean-Field Theory)	Macro Base Station (MBS) coverage, Small Base Station (SBS) density, IoT, transmission power, energy budget.
[14]	URLLC and eMBB	Coexistence, scheduling, multiplexing, and resource allocation with optimization.	Article proposes URLLC multiplexing with energy optimization and greedy algorithm.	BER, latency, energy consumption, resource block size, response time, energy efficiency, and data rate for eMBB.
[15]	URLLC and eMBB	Coexistence and resource allocation (main focus), packet scheduling (secondary focus).	The article proposes a Q-learning-based algorithm known as Latency-Reliability-Throughput Improvement in 5G NR using Q-Learning (LRT-Q).	Latency, reliability, throughput, and convergence time of Q-learning-based algorithms.
[16]	URLLC and eMBB	Coexistence, packet scheduling, and resource allocation.	Heuristic algorithm and unilateral matching game.	Minimum Expected Achieved Rate (MEAR) and fairness.
[17]	URLLC and eMBB	Resource allocation (main focus), traffic scheduling, and coexistence.	Deep Reinforcement Learning (DRL) using the Proximal Policy Optimization (PPO) algorithm.	Average reward, percentage of eMBB codewords in outage, average number of remaining URLLC packets in queue, latency, and comparative performance.
[18]	URLLC and eMBB	Optimization to address service multiplexing, ensuring both traffic types coexist without performance degradation. Traffic scheduling and resource allocation are directly addressed.	Combines Decomposition-Relaxation-Optimization Algorithm (DROA) and Twin Delayed Deep Deterministic Policy Gradient (TD3) for resource allocation and scheduling of eMBB and URLLC.	Average data rate of eMBB users, Service Level Agreement Satisfaction Ratio (SSR), fairness index, UAV energy consumption, Personalized Performance Fluctuation (PPF), learning efficiency of the proposed algorithm.
[19]	URLLC and eMBB	Formulated a non-convex optimization problem for 5G service coexistence, ensuring QoS.	Proposes: (1) a Hybrid orthogonal/non-orthogonal Multiple Access (HMA) and (2) a two-step algorithm based on Particle Swarm Optimization (PSO) to determine optimal transmission power values.	Number of supported URLLC UEs, eMBB UEs data transmission rate, URLLC transmission suMachine Learningccess probability, and QoS.
[20]	eMBB and mMTC	Addresses eMBB-mMTC coexistence in 5G, optimizing resources and reducing Random Access Channel (RACH) congestion in mMTC through NOMA plus PPO.	Proposes a PPO-DRL solution for eMBB-mMTC coexistence. Uses Non-Orthogonal Multiple Access (NOMA) with Successive Interference Cancellation (SIC) to: (i) manage eMBB-mMTC overlap, (ii) separate signals, increasing spectral efficiency.	Data transmission rate, percentage of eMBB in outage, convergence process of the proposed algorithm.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[21]	URLLC and eMBB	Studies URLLC-eMBB co-scheduling using puncturing in Multiple-Input Multiple-Output (MIMO) and NOMA for spectrum sharing, considering distinct service requirements.	Applies Gale-Shapley (GS) theory for user selection and Successive Convex Approximation (SCA) for energy allocation, proposing a low-complexity iterative algorithm and puncturing in MIMO-NOMA for coexistence scheduling.	Throughput, latency, reception success rate, user fairness index, and computational complexity.
[22]	eMBB, URLLC, and mMTC	Addresses the challenge of heterogeneous traffic coexistence in smart factories, focusing on resource management. Formulates a max-min optimization problem integrating task scheduling, bandwidth allocation, and robotic trajectory definition.	Proposes Task Scheduling, Bandwidth Allocation, and Robot Trajectory (TSBART), an algorithm optimizing task scheduling, bandwidth allocation, and robotic trajectory for better resource management in heterogeneous traffic.	Average Energy Efficiency (EE) of mMTC Devices, Minimum Average Spectrum Efficiency (SE) of the Robot, Probability of Satisfying Instantaneous Rate Requirements, and Algorithm Convergence Behavior.
[23]	URLLC and eMBB	The central problem lies in efficient resource allocation in service coexistence scenarios, aiming to: (i) maximize eMBB throughput, and (ii) ensure QoS requirements for URLLC.	Proposes a hybrid approach integrating: (i) contract theory, through an overlay/puncturing scheme, and (ii) matching theory to solve resource allocation problems in URLLC and eMBB coexistence scenarios.	Transfer rate, base station profit, reliability, latency, and comparison between schemes.
[24]	URLLC and eMBB	Proposes coordinated 5G resource allocation for eMBB-URLLC coexistence, maximizing eMBB's MEAR without affecting URLLC QoS.	Hybrid approach combining overlay and NOMA to improve spectral efficiency, using puncturing in mini-slots for urgent URLLC packets. Matching theory ensures fair resource allocation and QoS, while a low-complexity resource allocation algorithm maximizes MEAR.	Average eMBB data rate, MEAR, and Jain fairness index.
[25]	URLLC and eMBB	Develops a resource allocation issue for the coexistence of eMBB and URLLC via non-convex optimization.	Adopts a model based on Genetic Algorithms (GA), specifically the Data-Driven Genetic Algorithm-Based Spectrum Partition (DDGSP), to optimize spectrum allocation between URLLC and eMBB services.	Throughput, error rate (evaluates each method's ability to ensure URLLC reliability), and computational complexity.
[26]	URLLC and eMBB	Formulates a mini-slot optimization problem to maximize data rate, QoS, and eMBB stability and reduce resource losses caused by URLLC.	Proposes a dynamic resource allocation scheme based on DRL. Implementation uses an advanced Deep Q-Network (DQN) algorithm, operating at the mini-slot level to optimize spectral efficiency.	Comprehensive Loss (quantifies the negative impact of puncturing on eMBB users), defined to maximize data rate, QoS satisfaction, and minimize eMBB user data rate instability.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[27]	URLLC and eMBB	Addresses the critical challenge of efficient allocation of physical resources (spectrum, power) between two types of traffic with opposing requirements: URLLC and eMBB.	Approach based on DRL, specifically the Proximal PPO algorithm, to optimize resource allocation.	Total Episode Reward, percentage of eMBB codewords in outage, latency, and reliability.
[28]	URLLC and eMBB	Formulates a joint resource scheduling problem (frequency and power) for eMBB and URLLC traffic in Multi-Connectivity (MC) scenarios. It is a Mixed Integer Nonlinear Programming (MINLP) problem.	Presents an MC-based approach, modified effective capacity model, network slicing, Traffic Steering (TS), and two-step optimization.	Throughput, latency, queue length, resource utilization efficiency, ratio of associated UEs to UEs in MC.
[29]	eMBB, URLLC, and mMTC	Formulates a dynamic radio resource allocation problem for the Mobile Network Operator (MNO) to multiple Mobile Virtual Network Operators (MVNOs), ensuring coexistence of the three 5G service categories.	Proposes a multi-tenant slicing approach for the RAN, integrating: (i) dynamic scheduling mechanisms, and (ii) game theory-based models. Additionally, introduces an analytical model based on queueing theory.	Throughput, URLLC dwell time, URLLC queue waiting time, mMTC blocking probability, resource allocation, and resource utilization per operator.
[30]	URLLC and eMBB	Addresses the challenge of resource allocation in the coexistence of eMBB and URLLC services in 5G networks, emphasizing dynamic optimization of radio resources. Formulates a MINLP problem classified as NP-hard.	Puncturing mechanism, Quality of Experience (QoE)-aware Utility Function, iterative algorithm, heuristic resource allocation algorithm, and URLLC puncturing algorithm.	Number of URLLC users and their impact, average system utility, eMBB throughput, latency, and reliability.
[31]	URLLC and eMBB	Addresses the challenge of coexistence between URLLC and eMBB services in the same spectrum. Formulates a Multi-Objective Optimization (MOO) problem subject to QoS constraints.	Proposes a dynamic multiplexing approach based on Preemption Indication (PI) in the uplink and overlay through improved power control in the uplink.	URLLC capacity (maximum packet arrival rate), Resource Utilization Efficiency, Block Error Rate (BLER), latency, and reliability.
[32]	URLLC and eMBB	Addresses the challenge of coexistence and efficient traffic scheduling for eMBB and URLLC service users in 5G NR networks, especially given URLLC packet latency requirements.	Greedy algorithm based on queueing theory.	Throughput, reliability, latency, distribution of punched resource blocks (validates the mechanism's effectiveness in minimizing eMBB impact), URLLC service outage probability.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[33]	URLLC and eMBB	Addresses the challenge of enabling coexistence of eMBB and URLLC services in modern communication systems.	Proposes an integrated architecture combining: (i) Multi-access Edge Computing (MEC), (ii) Network Function Virtualization (NFV), (iii) dynamic allocation of virtual resources, and (iv) Continuous Time Markov Chain (CTMC).	Availability (system's ability to offer the minimum amount of functional and accessible virtualized network functions) and response time (interval between service arrival at the MEC-NFV node).
[34]	URLLC and eMBB	Analyzes the coexistence of eMBB and URLLC services in 5G-Advanced/6G networks, formulating a MOO problem to minimize E2E energy consumption and resource allocation costs while ensuring QoS requirements.	Introduces the Joint Radio and Core Resource Allocation (JRCRA) iterative algorithm, a scheme that coordinates: (i) spectrum and power allocation in the RAN, and (ii) computational resource management in the Core Network (CN). The solution is based on MINLP.	E2E energy consumption, resource usage cost, E2E latency, throughput, comparative performance.
[35]	URLLC and mMTC	Addresses the multi-objective challenge of simultaneously optimizing: (i) EE in NB-IoT systems and (ii) latency in critical (URLLC) and massive (mMTC) services.	Proposes four suboptimal algorithms: (i) heuristic, (ii) modified Shortest Job First (SJF), (iii) score-based algorithm, and (iv) multidimensional algorithm.	Energy efficiency, latency, data rate, number of repetitions, Signal-to-Interference-plus-Noise Ratio (SINR), and transmission power.
[36]	URLLC and eMBB	Addresses the resource allocation problem to multiplex eMBB and URLLC services in a 5G network. The problem is formulated as a non-convex optimization problem.	Proposes a hybrid puncturing and overlay scheme based on DRL. The approach uses the PPO algorithm to solve the non-convex optimization problem.	eMBB data transmission rate, probability of failed eMBB codewords, URLLC reliability, and latency.
[37]	URLLC and eMBB	Analyzes dynamic resource allocation and service scheduling in 5G networks, formulating an NP-hard and non-convex optimization problem.	Proposes a hybrid architecture based on DRL, with a specific approach using Deep Double Q-Learning (DDQN), integrating Thompson Sampling and puncturing techniques.	Throughput, reliability, latency, algorithm convergence.
[38]	eMBB, URLLC, and mMTC	Addresses the coexistence of eMBB, URLLC, and mMTC services in the downlink of a 5G NR network. Formulates a MINLP problem, inherently non-convex.	Primarily uses a DRL approach with the PPO algorithm to optimize resource allocation. Successive Convex Approximation (SCA) is employed as the basis for evaluating the proposed DRL performance.	Achievable data rate loss for eMBB, number of Mini Resource Blocks (mRBs) in eMBB outage, comparative performance with reference algorithms, and computational complexity.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[39]	URLLC and eMBB	Addresses energy-efficient resource allocation in heterogeneous cloud Radio Access Networks (H-CRAN). Formulates a non-convex MINLP problem.	Proposes an iterative algorithm combining integer relaxation, Big-M formulation, Dinkelbach method, auxiliary variable approximation, SCA, and a hybrid scheduling scheme Hybrid OMA/Hybrid NOMA (H-OMA/H-NOMA).	Energy efficiency, throughput, algorithm convergence.
[40]	URLLC and eMBB	Proposes a convex optimization model considering eMBB users' error correction capacity, aiming to maximize their sum rates while ensuring Physical Resource Block (PRB) scheduling and immediate URLLC traffic handling.	Main method used is a heuristic joint resource scheduling scheme for eMBB and URLLC traffic.	Throughput, eMBB rate gain, PRB allocation, URLLC traffic, and impact of URLLC traffic arrival rate.
[41]	URLLC and eMBB	Aims to minimize eMBB rate loss caused by overlay and puncturing to accommodate URLLC traffic. Formulates a MINLP problem.	Adopts a low-complexity resource allocation scheme for a base station serving both services (URLLC and eMBB) in the downlink.	Validates eMBB data rate loss and URLLC reliability, temporal complexity, traffic load, packet size, and channel conditions.
[42]	URLLC and eMBB	Investigates the coexistence of eMBB and URLLC services in 6G networks, formulating a MINLP problem to optimize URLLC packet acceptance while minimizing eMBB rate impact.	Proposes a bipartite matching approach, NOMA, overlay or puncturing techniques, and the GS algorithm to optimize resource allocation.	URLLC packet admission rate (proportion of URLLC packets admitted compared to total arriving packets), eMBB data rate loss, and URLLC reliability.
[43]	URLLC and eMBB	Explores the coexistence of eMBB and URLLC services in 5G networks, emphasizing optimization of flexible and self-adaptive Transmission Time Interval (TTI) intervals for each service.	Proposes the Self-Adaptive Flexible Transmission Time Interval Scheduling (SAFE-TS) strategy, based on machine learning.	Delay, packet loss rate (measures URLLC service reliability), classification accuracy, throughput.
[44]	eMBB and mMTC	Proposes optimizing resource allocation for the coexistence of eMBB and mMTC services in the same RAN.	Utilizes network slicing and NOMA techniques to optimize multi-service performance in a RAN.	Throughput, maximum number of connected devices, reliability, average channel gains.
[45]	eMBB, URLLC, and mMTC	Explores efficient resource allocation in 5G NR networks to optimize eMBB, URLLC, and mMTC performance, evaluating relationships between power and service arrival rates.	Proposes using Rate-Splitting Multiple Access (RSMA) to optimize resource allocation in service coexistence scenarios in 5G networks.	Throughput, latency, reliability, maximum number of connected mMTC devices.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[46]	URLLC and eMBB	Addresses resource allocation optimization in 5G networks for different traffic types.	Proposes a hybrid NOMA method combining Near-Far/Far-Near (NF-FN) and Near-Near/Far-Far (NN-FF) user pairing strategies to optimize resource allocation in 5G networks.	Spectral efficiency, throughput, latency, fairness index, and outage probability.
[47]	URLLC and eMBB	Investigates eMBB and URLLC coexistence in 5G/B5G networks, formulating a MOO problem and converting it into Single-Objective Optimization (SOO) using the Objective Product Method (OPM).	Proposes preemptive scheduling with unequal mini-slots and an Optimized Sparrow Search Algorithm (OSSPA) to improve coexistence between eMBB and URLLC services in 5G/B5G networks.	Number of supported URLLC users, eMBB throughput, resource utilization efficiency, user satisfaction.
[48]	URLLC and eMBB	Analyzes URLLC and eMBB traffic coexistence in 5G NR networks, proposing an optimization problem to balance resource allocation and minimize session losses or preemptions.	Explores various resource allocation approaches and AI/ML techniques, using a multidimensional Markov chain-based queueing model to define resource needs.	Session drop probability, session preemption probability, system resource utilization.
[49]	URLLC and eMBB	Addresses resource scheduling for eMBB and URLLC service coexistence, proposing efficient spectrum resource allocation to minimize eMBB performance losses and meet URLLC latency requirements.	Proposes a dual-dimension scheduling scheme with puncturing prediction for URLLC, using the TD3 algorithm.	Throughput, number of URLLC puncturing instances, URLLC resource occupancy rate, fairness index.
[50]	URLLC and eMBB	Addresses URLLC and eMBB coexistence in B5G networks, formulating two optimization problems to maximize fairness and minimize eMBB rate loss due to URLLC puncturing.	Combines matching theory with deep learning techniques, both supervised and unsupervised, to optimize resource allocation in 5G networks.	Fairness, throughput, reliability.
[5]	URLLC and eMBB	The study analyzes the state of the art of 5G, with an emphasis on the coexistence mechanisms between eMBB and URLLC traffic.	The study adopts the PRISMA statement to classify the reviewed works into five main categories: multiplexing, QoS, machine learning, network slicing, and C-RAN architecture. The analysis focuses on the contributions of each category to the coexistence of services in 5G networks.	binary throughput, latency, and reliability URLLC, fairness, spectral efficiency, energy efficiency, QoS.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[51]	URLLC and eMBB.	The study investigates the coexistence of URLLC and eMBB services in B5G networks, addressing the challenge of maximizing the sum rate of URLLC users while managing interference with eMBB services.	The research compares RSMA with Orthogonal Multiple Access (OMA) and NOMA in network slicing scenarios for URLLC and eMBB, highlighting the efficiency of RSMA in resource management and meeting QoS requirements.	Sum rate of URLLC and eMBB, reliability, latency
[8]	URLLC and eMBB.	The study focuses on the efficient allocation of resources in coexistence scenarios between eMBB and URLLC in 5G networks, proposing low-complexity solutions for NP-hard max-min problems.	The e-GREEDY algorithm is proposed, a low-complexity greedy approach.	Minimum Data Rate and Fairness Index
[52]	URLLC and eMBB.	This investigation analyzes the simultaneous operation of eMBB and URLLC in cellular networks, proposing a joint and stochastic optimization problem to ensure QoS requirements.	The authors propose a Block Coordinate Descent (BCD) algorithm and introduce the Dynamic Resource Allocation and Puncturing Strategy (DRAPS).	Average queue backlog of eMBB terminals, number, number of allocated RBs, and the probability of packet transmission of URLLC terminals.
[53]	URLLC and eMBB.	The study examines the coexistence of eMBB and URLLC in the MEC scenario. An MINLP optimization problem is formulated.	The research adopts a decomposition approach to the optimization problem into convex subproblems, followed by an iterative method to obtain an approximately optimal solution.	Although the authors consider data rate, delay, and reliability in the study, the simulation results indicate energy consumption as the only primary KPI used to validate the proposal and compare the proposed method with traditional approaches.
[54]	URLLC and eMBB.	The study addresses the complexity of joint resource scheduling for eMBB and URLLC, formulating the MINLP problem.	The study proposes a resource allocation strategy that considers the risks of violating URLLC delay specifications, applying the Conditional Value-at-Risk (CVaR).	Fairness, binary debt, algorithm convergence, and latency.
[55]	URLLC and eMBB.	The research focuses on the efficient multiplexing of eMBB and URLLC traffic, using grant-free allocation in the uplink of 5G NR networks.	The study involved detailed simulations to evaluate the impact of different power control configurations, applied to users of eMBB and URLLC trends.	The probability of URLLC failure assesses the dependability and latency of the URLLC service, while the 5th and 50th percentiles of the eMBB SINR gauge the effect on eMBB service capacity.

Table 2. *Cont.*

Ref.	Type of Service	Problem Formulation	Methods/Approaches	KPIs
[56]	mMTC and URLLC.	The study addresses the difficulty of simultaneously supporting the stringent URLLC requirements in critical mMTC scenarios, considering the limited radio resources in future wireless networks beyond 5G.	The study reviews technologies for mMTC and URLLC, identifies challenges arising from conflicting requirements, and explores potential solutions for critical mMTC across various layers of the network.	The article does not present a proposal with results validated by KPIs, but highlights relevant KPIs for mMTC (massive link density, maximum coupling loss, and battery life) and for URLLC (BLER and latency) separately.
[57]	URLLC and eMBB.	The study addresses the coexistence of eMBB and URLLC in 5G networks. An NP-hard optimization problem is formulated due to its MINLP nature.	The study proposes a hybrid solution that combines a heuristic algorithm with an approach based on Graph Neural Networks (GNNs).	Fairness, binary throughput, Packet Loss Ratio (PLR).
[58]	URLLC and eMBB.	The study addresses the challenge of ensuring that the stringent URLLC standards are met while the services of this trend coexist with eMBB traffic in the unlicensed spectrum.	The study proposes an opportunistic preemptive approach and explores the calibration of maximum sizes of contention windows for URLLC and eMBB.	Reliability of URLLC, latency of URLLC, and transmission rate of eMBB.

3.1. Co-Occurrence Network of Keywords

Figure 2 shows the keyword co-occurrence network generated in VOSviewer version 1.6.20 with a minimum threshold of six occurrences. The diagram reveals the conceptual structure of the most frequent topics in the corpus, identifies current research trends, and highlights the relationships among the principal themes addressed in the literature. The network comprises three clusters, each depicted in a distinct color (red, green, and blue).

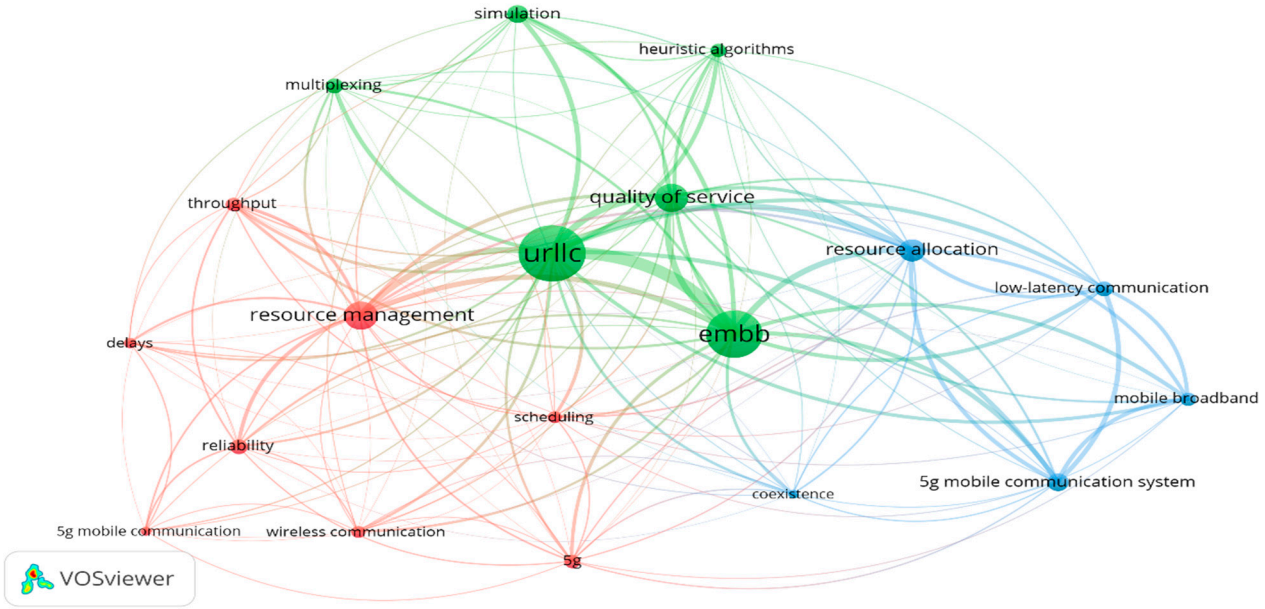

Figure 2. Co-occurrence network of keywords.

Red cluster (eight items): The hub term is resource management, and the associated keywords include 5G, 5G mobile communication, scheduling, wireless communication, delays, reliability, and throughput. This cluster emphasizes efficient resource management and the challenges in ensuring the high reliability and performance of 5G wireless networks. Scheduling and wireless communication denote resource allocation strategies and connectivity maintenance, respectively, whereas delays, reliability, and throughput represent critical performance metrics.

Green cluster (six items): Anchored by URLLC, eMBB, and QoS, this cluster addresses QoS optimization through algorithmic techniques. Supporting terms such as heuristic algorithms and multiplexing refer to methods employed to satisfy QoS requirements, whereas simulation indicates the validation tools used to evaluate performance improvements.

Blue cluster (five items): This cluster highlights strategies for efficient resource distribution in mobile networks to enable latency-sensitive applications. Related terms include coexistence, mobile broadband, and 5G mobile communication systems, underscoring the need to balance diverse traffic types within the shared radio resources.

The keyword co-occurrence network underscores the interrelated nature of these challenges and indicates that current research emphasizes intelligent resource-management solutions, advanced allocation algorithms, and QoS-optimization strategies.

3.2. Most Cited and Viewed Articles

It is important to emphasize that the initial selection of articles for inclusion in this scoping review was conducted strictly according to the predefined thematic criteria (Section 2.1), entirely independent of any impact metrics. Consequently, Table 3 does not represent the inclusion criteria; instead, it offers an additional interpretative layer applied to the already curated corpus.

Table 3. Most Cited and Viewed Articles.

No.	Ref.	Citations	Views
1	[56]	129	7064
2	[43]	57	4527
3	[24]	42	2351
4	[15]	41	1875
5	[45]	38	1697
6	[16]	37	-
7	[23]	35	1748
8	[28]	31	1581
9	[52]	27	1872
10	[54]	27	1489
11	[51]	23	2194
12	[41]	21	1800
13	[31]	19	1117
14	[21]	18	1394
15	[55]	17	950

Table 3. *Cont.*

No.	Ref.	Citations	Views
16	[44]	16	873
17	[27]	15	790
18	[5]	14	38
19	[37]	11	1864
20	[35]	9	1680

The articles are ordered according to a hybrid metric that combines citation counts and views, which serve in this context as proxy indicators of consolidated academic impact and current attention from the research community, respectively. This approach supports the identification of both foundational works and emerging trends within the thematic domain under analysis. The listed studies address packet scheduling, spectrum allocation, AI-driven resource optimization, network slicing, and the coexistence of multiple service categories.

4. Discussion

4.1. Annual Distribution of Published Articles

The annual distribution of the identified articles, presented in Figure 3, reveals a steady increase in the volume of publications on service coexistence in 5G NR networks, with pronounced peaks in 2021 and 2024. This pattern is consistent with the progressive maturation of the research field in response to practical deployment challenges. The initial peak in 2021 coincides with the early stages of commercial 5G rollout, while the more significant peak in 2024 can be understood in the context of the growing maturity of critical applications that rely on the robust coexistence of heterogeneous services. The reviewed literature [22,29,35] consistently reports that the expansion of domains such as Industry 4.0 (featuring autonomous production lines and collaborative robotics), advances in telehealth (such as remote surgery and continuous patient monitoring), and the large-scale deployment of sensors for smart cities have intensified the demand for academic solutions to the scheduling and resource allocation problems discussed in this review.

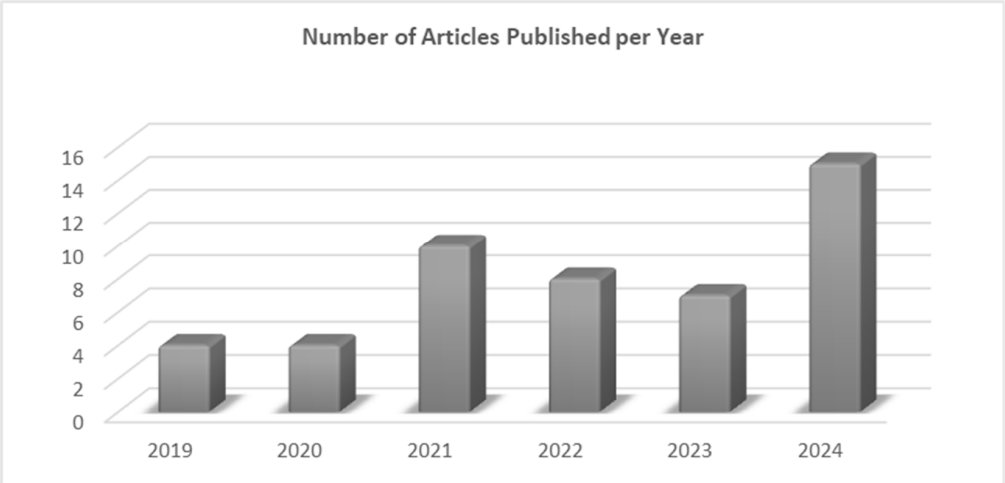

Figure 3. Annual distribution of identified scientific articles.

Regarding the distribution of the investigated service combinations (Figure 4), the clear predominance of studies focused on eMBB–URLLC coexistence (81.25%) relative to scenar-

ios involving mMTC (only 18.75%) reflects the technical challenges inherent to each pairing. The scarcity of works that integrate mMTC, particularly in combination with URLLC, is technically justifiable. As detailed by [56] and corroborated by the complexity of the optimization problems formulated in the included studies addressing this coexistence [35], reconciling the high device density and energy efficiency (fundamental to mMTC) with the ultra-low latency and extreme reliability requirements (essential to URLLC) constitutes a well-recognized challenge, often modeled as NP-hard and multi-objective.

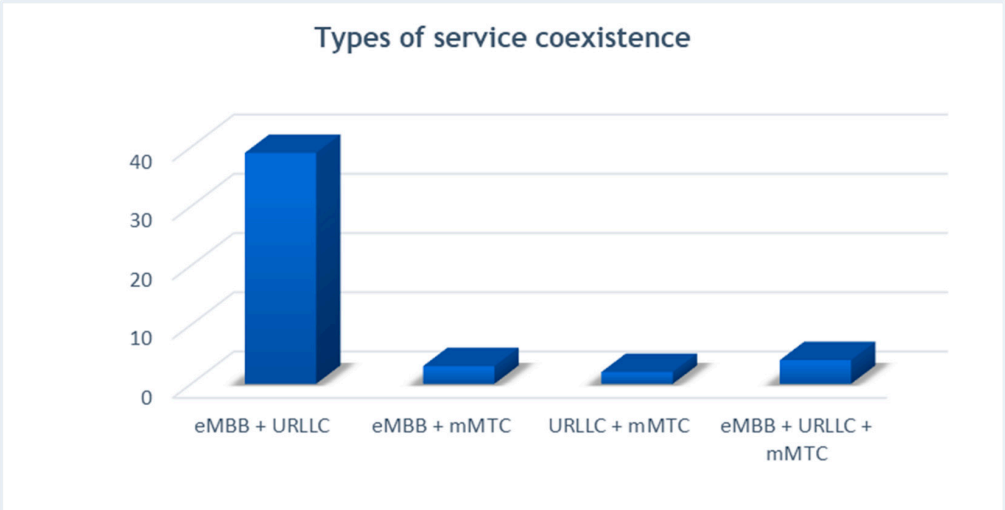

Figure 4. Main combinations between 5G services.

This intrinsic complexity, in contrast to the more established trade-offs between data rate and latency in the eMBB–URLLC pairing, has tended to direct initial research efforts toward more tractable problems, which helps explain the distribution observed in the literature.

4.2. Main Combinations Between 5G Services

Figure 4 shows the distribution of studies addressing various coexistence scenarios among the three primary service types in the 5G and 6G networks. Each analyzed combination reflected a distinct research focus. Notably, the coexistence of eMBB and URLLC has received the most attention from researchers and has demonstrated significant growth.

This trend suggests a high demand for service configurations that simultaneously require a high data throughput and strict reliability, or latency guarantees. The increased research interest can be attributed to emerging applications such as remote surgery involving real-time video transmission, which demands both ultra-reliable and high-bandwidth communications.

Research on URLLC–mMTC and eMBB–mMTC coexistence scenarios has expanded more slowly, suggesting that practical deployment is still at an early stage. eMBB–mMTC pairing appears to be less critical, perhaps because its requirements—high data-rate throughput versus massive device connectivity—are not inherently antagonistic. In contrast, the coexistence of URLLC and mMTC remains sparse, largely because of the difficulty in reconciling ultralow latency with a high node density.

The simultaneous coexistence of all three service categories (eMBB, URLLC, and mMTC) has received little attention, although interest has been increasing at a moderate pace. This tri-service scenario is the most demanding: eMBB requires substantial bandwidth; URLLC mandates absolute priority and reliability; and mMTC requires energy efficiency and scalability. Such a comprehensive coexistence is expected to become critical

in the forthcoming 6G systems, which must support all service classes concurrently in environments such as smart cities, intelligent buildings, and fully automated factories.

In summary, Figure 4 shows that most studies (approximately 81.25%) focus on situations where eMBB and URLLC occur simultaneously, whereas fewer studies (approximately 18.75%) look at mMTC scenarios. These quantitative results obtained from the analysis of the studies indicate that there is a need for immediate research on combinations involving mMTC, including triples (eMBB + URLLC + mMTC).

4.3. Main Methods Used

Figure 5 presents the primary methods employed to address challenges related to service coexistence, packet scheduling, and resource allocation in 5G networks. The selection criteria required each method to appear in at least three of the reviewed sources, thereby ensuring that only the most frequently cited and relevant approaches in the recent literature were included.

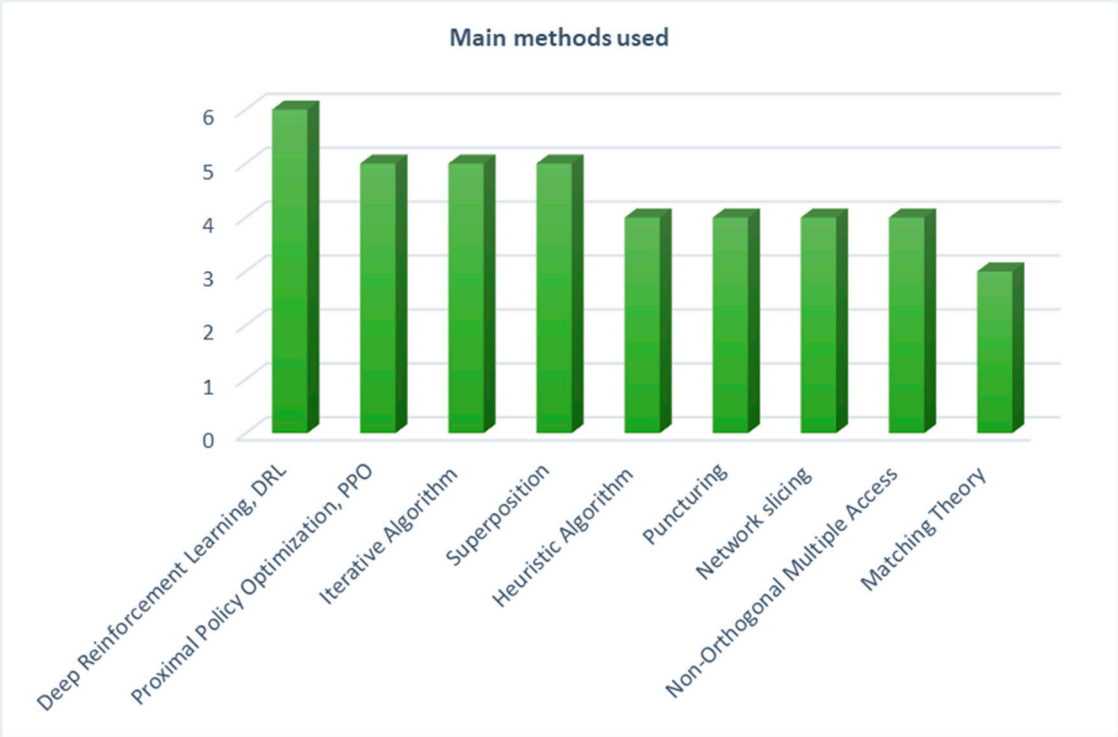

Figure 5. Main methods used.

The predominant approach is based on Artificial Intelligence (AI) techniques, specifically Reinforcement Learning, which confirms the current trend of applying AI to handle dynamic environments and to solve complex and uncertain problems in 5G and future networks. The use of hybrid schemes, combining classical and advanced methods, such as heuristic techniques and iterative approaches, remains common owing to their computational efficiency and ease of implementation.

On the other hand, techniques such as service overlapping, NOMA, and network slicing highlight the relevance of structural strategies that enable the coexistence of multiple services using the same physical resources with minimal interference.

Therefore, the graph shows that the solution for service coexistence in 5G and subsequent networks is strongly directed toward the integration of Artificial Intelligence and dynamic multiplexing, complemented by traditional heuristic techniques and structural approaches. Next, we describe these methods in detail.

4.3.1. Puncturing

Puncturing is a fundamental technique proposed by 3GPP to enable the coexistence of different types of services within the same radio spectrum [21,31]. This technique is used to optimize resource allocation and improve spectral efficiency in scenarios with heterogeneous traffic. It allows the simultaneous transmission of different types of data by dynamically adjusting network resources to meet the specific requirements of each service.

Frequently referred to in the literature as an effective solution for the coexistence of eMBB and URLLC services, this technique is based on the temporary interruption of eMBB traffic transmission on certain network resources (such as mini-slots within a full slot) to accommodate the immediate transmission of URLLC traffic [23,50,59]. The latter presents stricter requirements, namely, ultralow latency and high reliability, in addition to a sporadic and unpredictable nature.

Puncturing is considered an effective approach to increase spectral efficiency, allowing URLLC traffic to use resources previously allocated to eMBB, ensuring its immediate scheduling. However, this technique inevitably compromises the performance of eMBB.

The challenges associated with preemption consist of simultaneously balancing latency, reliability, and spectral efficiency, while managing the inherent complexity of dynamic resource allocation. Among the main challenges are:

- Conflicts of requirements between different types of service (trade-off) [2,21,60,61];
- Allocation of resources across multiple time scales [27];
- Reduction in transmission rate and reliability of eMBB services, resulting from interruptions or preemption by URLLC traffic [2,21,60];
- High computational complexity associated with the management and scheduling process [2,30].

To address the balance between strictly meeting URLLC requirements and mitigating the degradation of eMBB QoS caused by puncturing, several studies have focused on optimizing puncturing schemes that balance the performance of both services [8,16,30,52].

Among the proposed approaches, DRL is used for dynamic resource slicing decisions with puncturing [26,27,37]; game theory-based algorithms (matching theory) for efficient resource allocation and service matching in puncturing scenarios [50]; and the development of analytical models that quantify and mitigate the eMBB rate loss caused by the preemption of URLLC traffic.

Some proposals have also explored the combination of puncturing with signal superposition techniques (superposition coding) to mitigate the negative impacts on the eMBB performance [23,42]. The granularity of puncturing (for example, at the mini-slot level) and the dynamics of URLLC traffic arrival thus prove to be determining factors in the design of efficient schemes and should be carefully considered in the modeling and implementation of realistic and effective solutions [40]. Strategic solutions in artificial intelligence and optimization algorithms present innovative advancements.

4.3.2. Superposition

The Superposition technique represents a fundamental approach to solving coexistence challenges by allowing URLLC traffic to be overlaid on eMBB traffic in mini-temporal windows, reducing URLLC latency without significantly compromising eMBB data rates.

This technique includes efficient resource allocation with the aim of ensuring that both services meet their respective QoS requirements as well as dynamic scheduling, which guarantees that URLLC traffic is processed proactively and reactively, achieving an ideal balance between low latency and high reliability.

Its importance lies in its ability to allow different types of traffic to share the same radio resources in a non-orthogonal manner with the aim of improving spectral efficiency

and accommodating various QoS requirements. Superposition allows signals from multiple services to be transmitted simultaneously at the same frequency and time. The potential to increase spectral efficiency compared with orthogonal allocation approaches, in which resources are exclusively divided among services, constitutes one of the main advantages of this technique. By allowing an overlap, it is possible to support more users and services with the same available spectrum.

This technique can be implemented in different ways:

- Superposition in the power domain occurs when signals are combined with different power levels according to the QoS requirements of users;
- Superposition in the time domain occurs when different time intervals or subframes are allocated to eMBB and URLLC, ensuring coexistence without direct interference.

Despite allowing the simultaneous transmission of signals from various services over the same frequency and temporal resources, the overlay technique continues to face significant challenges, including the efficient allocation of resources, the assurance of conflicting QoS requirements, and high computational complexity.

For successful implementation, it is essential to rigorously manage interference, carry out efficient energy distribution, and develop precise models that ensure compliance with QoS requirements for both services.

The solutions proposed in areas such as game theory, reinforcement learning, and hybrid multiplexing schemes have contributed to overcoming the obstacles associated with the superposition technique, thereby revealing its potential in future communication networks. The main challenges associated with superposition in 5G networks and beyond are:

- Conflicting QoS [2,24,36];
- Interference between services [2,24,25,36,62];
- Non-convex resource allocation [2,36,38];
- Computational complexity [62,63].

4.3.3. Heuristic Algorithm

A heuristic algorithm represents a pragmatic method for addressing clearly defined mathematical challenges through intuitive reasoning, examining the inherent framework of the problem to deduce a satisfactory solution, although not always the ideal or most precise one [64,65]. These algorithms are distinguished by their rapidity, efficacy, and flexibility and serve a pivotal function in addressing intricate challenges. Their significance arises from their capacity to provide efficient solutions with minimal computational complexity for issues that would otherwise be insurmountable, owing to their NP-hard characteristics and real-time limitations intrinsic to wireless networks [16,30,57].

In the reviewed literature, their practicality is evidenced for specific coexistence challenges. For instance, the authors in [16] employed a matching theory-based heuristic to efficiently associate eMBB users with URLLC requirements, significantly improving the MEAR and fairness. Similarly, the authors in [8] developed a low-complexity greedy heuristic (e-GREEDY) for real-time eMBB-URLLC scheduling, achieving performance close to an optimal solver.

However, although heuristic algorithms are indispensable for making the complex coexistence problems in 5G solvable in real time, they present inherent challenges, such as:

- Renunciation of global optimality in favor of speed and efficiency [14,16,57];
- Increased complexity and computational cost in more robust variants (e.g., metaheuristics) [25,26];

- Difficulty in ensuring strict QoS requirements under high loads or in heterogeneous scenarios [25,47];
- Performance trade-offs across multiple metrics [5].

4.3.4. Network Slicing

Network slicing consists of the practice of dividing a physical network into several isolated logical networks called slices, which can be customized according to different service requirements [66]. This approach is enabled by advances in NFV and Software-Defined Networking (SDN), which decouple hardware from software functions and allow dynamic, programmable resource management.

This technology is paramount for mitigating the challenges associated with coexistence, resource distribution, and service orchestration within 5G networks. It facilitates the provision of a wide array of services while guaranteeing service isolation, preserving QoS, and enabling adaptable and effective resource allocation. For example, the authors of [29] proposed a multi-tenant slicing approach for the RAN, integrating dynamic scheduling mechanisms, models based on game theory and queue theory, to significantly improve QoS parameters, ensure effective isolation between different MVNOs, and ensure user satisfaction across all network slices. Similarly, the authors [45] integrated RSMA within network slices to optimize resource allocation for heterogeneous traffic, achieving superior spectral efficiency compared to OMA and NOMA.

However, despite its benefits, the effective implementation of network slicing encounters numerous challenges, including:

- Intricacies of managing and orchestrating multiple slices [5];
- Inter-slice interference [5,29];
- Security and isolation [66];
- Resource allocation efficiency [29,44].

Consequently, continuous scholarly inquiry into areas such as AI-driven slicing algorithms, standardized interfaces for RAN slicing, and lightweight security protocols to address these challenges and fully realize the potential of network slicing in 5G and beyond.

4.3.5. NOMA

NOMA is an innovative radio access technology designed for 5G networks and future generations of mobile communication. Unlike conventional Orthogonal Multiple Access (OMA) techniques, NOMA allows multiple users to share the same resource block at different power levels, thereby increasing the spectral efficiency of the system [44,46,67].

This technology is notable for its substantial enhancements in spectral efficiency, connectivity, equity, support for heterogeneous service coexistence, resource allocation flexibility, and integration with other advanced technologies such as MIMO, network slicing, beamforming, puncturing, and superposition. For example, the authors in [20] combine NOMA with SIC to resolve the overlap between eMBB and mMTC, therefore separating the signals to improve spectral efficiency. A hybrid technique is presented in [24] that combines superposition and NOMA to improve spectral efficiency, yielding a better MEAR compared to methods based on contract theory and puncturing. The authors in [46] use network slicing and NOMA techniques to improve the performance of diverse services in a RAN. The authors of [42] propose a methodology using bipartite matching theory, NOMA, overlay or puncturing techniques, and the GS algorithm to improve resource allocation.

Despite its benefits, NOMA implementation presents several challenges, as discussed in [46,68–70]:

- Complexity in the receiver due to the need for techniques such as SIC;
- Intercell interference;

- Sensitivity to channel conditions;
- Power distribution.

Consequently, contemporary research underscores the advancement of robust SIC designs, AI-enhanced power distribution, and the integration of alternative techniques such as beamforming to tackle these challenges and optimize NOMA's potential as an essential technology for the coexistence of diverse services in 5G and future communication systems.

4.3.6. Matching Theory

Matching theory provides a mathematical framework for establishing efficient and stable associations between two sets of entities (e.g., users–resources, services–channels) within 5G/6G networks, optimizing resource utilization and overall network efficiency under specific constraints [2]. This approach applies to scenarios such as cache peering, Device-to-Device (D2D) communication, spectrum sharing, and service-to-resource mapping.

In service coexistence, matching theory is applied to design fairness-aware allocation strategies. For example, the authors [24] proposed a matching theory-based algorithm to coordinate eMBB–URLLC resource allocation, maximizing MEAR while guaranteeing URLLC latency and reliability. Similarly, the authors [16] applied a one-to-many matching game to associate eMBB users with URLLC requirements, significantly improving the MEAR and fairness. Some approaches integrate matching theory with other techniques, such as puncturing [50], to further enhance spectral efficiency.

Although matching theory offers powerful tools for modeling and solving complex coexistence and resource management challenges in 5G networks, its practical application requires overcoming significant obstacles. Their main challenges are:

- Model complexity [2];
- Computational overhead [24];
- Dynamic adaptability under real-time network conditions [50];
- Integration with legacy systems: as highlighted in studies that emphasize low-complexity deployment and spectrum sharing in hybrid 4G/5G environments, the combination of strategies based on matching with existing infrastructure (e.g., LTE-A Pro) may require a protocol redesign and backward-compatible solutions [5,50,56].

Thus, further research in this area is necessary to develop more sophisticated and adaptable matching approaches that can be implemented efficiently in future wireless networks.

4.3.7. Iterative Algorithm

An iterative algorithm is a problem-solving method that employs a repetitive process, in which a sequence of steps or operations is executed multiple times until a stopping condition is met. During each iteration, the algorithm updates or refines a partial solution, seeking to progress toward a final or optimized solution to the problem.

In 5G NR networks, these algorithms play a significant role in solving complex problems classified as NP-hard or formulated as MINLP, which are difficult to solve optimally in polynomial time, particularly in large-scale and real-time network scenarios [21,38,41].

In the context of coexistence, iterative algorithms provide a practical means of obtaining suboptimal solutions or local optima while maintaining controlled computational complexity [5]. For example, the authors in [39] proposed an iterative algorithm to address a non-convex optimization problem in H-CRAN networks, combining auxiliary variables, Successive Convex Approximation (SCA), integer variable relaxation, Big-M formulation, and the Dinkelbach method. A hybrid scheduling scheme (H-OMA/H-NOMA) was also employed to balance energy efficiency and throughput. In [21], an iterative algorithm based on Difference Convex (DC) programming and SCA was applied to solve a non-convex

optimization problem. Both methods were used jointly to adjust power allocation and ensure convergence to a viable and efficient solution. Similarly, Wang et al. [30] employed an iterative decoupling optimization algorithm combined with a heuristic approach to manage the coexistence of eMBB and URLLC traffic, maximizing system utility while meeting latency and reliability requirements. This strategy alternated between optimizing one variable (resource allocation) and another (puncturing URLLC) until reaching an acceptable equilibrium solution.

Despite their relevance in dealing with complex problems and the variability of network conditions, iterative algorithms present inherent challenges and limitations, as discussed in [21,30,39]. Among these, the following stand out:

- Convergence and stability;
- Computational complexity;
- Sensitivity to initial parameters;
- Scalability;
- Generalization and adaptation in highly dynamic environments.

Thus, while iterative algorithms remain a vital tool for theoretical analysis and as a performance benchmark, their practical deployment often necessitates hybridization with AI-based methods to achieve both efficiency and optimality in 5G and beyond networks.

4.3.8. Proximal Policy Optimization

PPO is a DRL algorithm that has proven to be a powerful technique for solving complex problems related to coexistence, resource allocation, and service scheduling in 5G/6G networks [17,20,27].

It is a versatile method that can be used in various contexts, such as in MEC, to optimize task offloading and resource allocation and reduce consumption time [71]. In the context of heterogeneous networks, this method can ensure a dynamic and balanced distribution of resources between central and peripheral users, improving network coverage and spectral efficiency and ensuring the satisfaction of QoS requirements [27,72].

The reviewed literature demonstrates the effectiveness of PPO in various scenarios. For example, the authors in [20] combine PPO and NOMA to optimize resources and reduce RACH congestion in mMTC. This approach minimizes interruptions in eMBB while accommodating URLLC devices, using dynamic scheduling to improve throughput and analyze overlap in congested scenarios. In [27], the method is applied to optimize resources, demonstrating superior performance compared to other traffic scheduling approaches, especially in minimizing outages without compromising QoS, and ensuring efficiency in terms of computational complexity. In [36], it is used in resource allocation to multiplex eMBB and URLLC services in a 5G network, aiming to minimize data loss for eMBB users due to coexistence with URLLC users. Similarly, in [38], it is also applied to optimize resource allocation, concluding that PPO is an efficient solution for multiplexing eMBB, URLLC, and mMTC traffic, ensuring balance among them.

Despite its advantages, the application of PPO in 5G faces challenges, as discussed in [20,27,38]. Among these, the following stand out:

- Difficulty training in complex environments;
- Dependence on an adequate reward function;
- Computational overload;
- Generalization and stability in highly dynamic environments.

Thus, PPO is a powerful and versatile tool for intelligent network management. Its reinforcement-learning-based structure allows for the continuous improvement of allocation strategies, consolidating it as a promising solution for intelligent resource management in next-generation wireless networks.

4.3.9. DRL

Deep Reinforcement Learning (DRL) plays a central role in addressing the challenges associated with coexistence, resource allocation, and service scheduling in 5G/6G networks. Its relevance stems from its ability to handle heterogeneity, optimize multiple QoS parameters, perform dynamic adaptation, and outperform conventional optimization methods [73,74].

The reviewed literature showcases DRL's success through specific algorithms. For example, the authors in [26] use DQN in a dynamic slicing scheme for resource allocation, operating at the mini-slot level, to reduce resource losses caused by URLLC and maximize data rate, QoS, and eMBB stability. In [37], DDQN is integrated with the Thompson Sampling and puncturing techniques to solve service coexistence problems, providing significant improvements in terms of performance and efficiency in resource allocation in next-generation mobile networks. Other DRL implementations, such as PPO, have already been presented in Section 4.3.8. Therefore, in addition to the selected literature, in [75], a DRL agent was implemented for Medium Access Control (MAC) layer scheduling, demonstrating its agnosticism to numerology and its consistent outperformance of conventional methods across key indicators.

However, despite providing adaptive and efficient solutions for optimizing resource utilization in complex and dynamic environments, DRL still faces significant challenges. Besides those listed in Section 4.3.8, we can list two more: slow convergence and overestimation of Q values. In this context, future research should focus on improving intelligent optimization mechanisms, developing advanced network slicing techniques, and enhancing integration with emerging technologies to increase overall efficiency and adaptability in next-generation networks.

4.4. Key Performance Indicators Used for Validation of Results

Key Performance Indicators (KPIs) are essential for evaluating the efficiency and QoS provided, enabling network operators to monitor performance in real time and adjust their systems as needed. Figure 6 presents the main KPIs used to validate the results from at least three selected sources in this scoping review. These indicators include the following metrics:

- Throughput;
- Latency;
- Reliability;
- Fairness (Jain Index);
- Energy Efficiency (EE);
- Block Error Rate (BLER);
- Spectral Efficiency (SE).

According to the analyzed data, throughput and the combination of latency/reliability emerged as the most recurring and prioritized indicators—an expected result given that these KPIs are the most critical and potentially conflicting in-service coexistence scenarios.

The emphasis on throughput highlights the importance of maximizing data transmission rates as a central objective in 5G networks, particularly for eMBB-type services. However, the focus on latency and reliability reflects the critical need to ensure ultrafast and highly reliable communications, a fundamental requirement for URLLC services.

Fairness ensures equitable resource distribution among competing users or services in heterogeneous environments. Meanwhile, EE reveals a growing concern for network energy consumption, especially in the face of massive connectivity demands (mMTC) and environmental sustainability imperatives.

The relatively less frequent presence of BLER and spectral efficiency suggests that these metrics, although technically relevant, assume less centrality in studies that aim to ensure efficient coexistence among multiple heterogeneous services. In the following section, we provide a detailed analysis of each KPI.

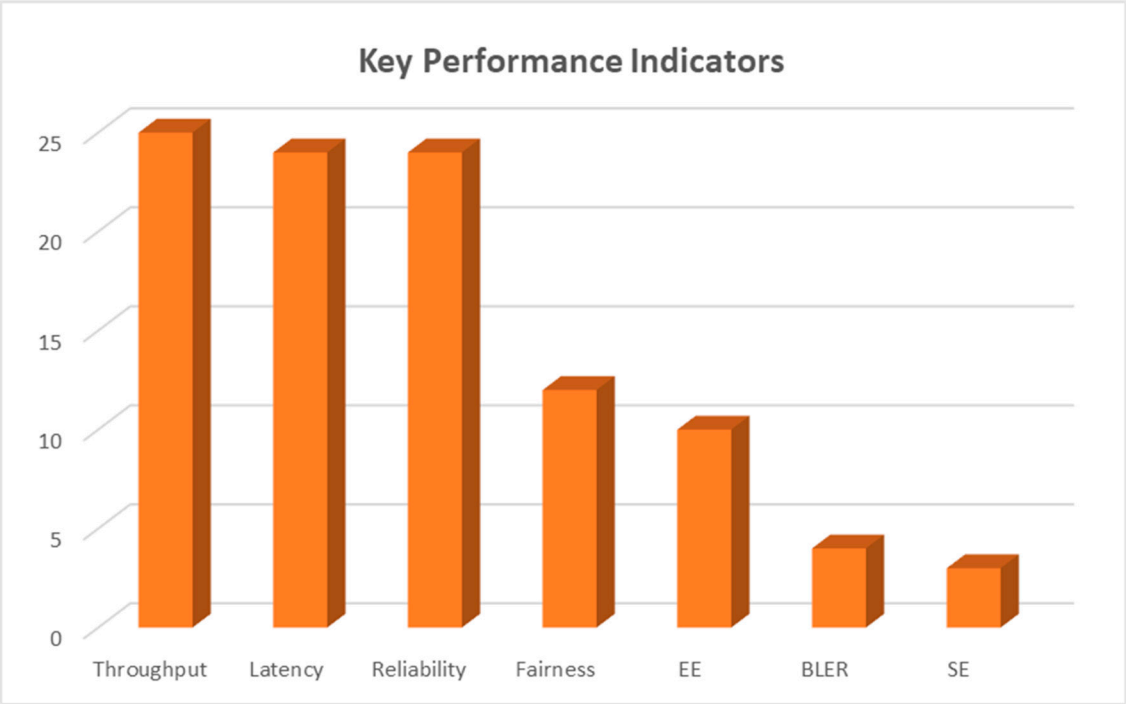

Figure 6. Main considered KPIs.

4.4.1. Throughput

Throughput is frequently considered to evaluate the performance of eMBB transmissions. This KPI appears under various designations in the analyzed sources, such as data rate [18,23,46], sum rate [36,37,46,51], achievable rate [45], and transmission rate [23,25,28,29,76].

It is a critical indicator in 5G networks, as it quantifies the efficiency of data transmission between devices and the network by measuring the number of successfully transmitted bits per unit of time. High throughput is essential for meeting the performance requirements of various applications and services enabled by 5G networks.

Its importance is evidenced by key factors, such as support for high data rates for demanding applications, such as high-definition video streaming, virtual reality, and augmented reality, as well as its relationship with complementary metrics, such as latency, spectral efficiency, QoS, and overall network performance evaluation.

In summary, throughput is a determining KPI in 5G networks, directly influencing the user experience in high-bandwidth applications, compliance with URLLC latency requirements, and degree of spectral efficiency. Thus, it plays a vital role in assessing and optimizing the network performance in multiple traffic scenarios.

4.4.2. Latency

Latency is a particularly relevant KPI in the context of URLLC services because of their stringent real-time communication requirements. This KPI is used both in the evaluation of scheduling methodologies [35,46] and in the analysis of essential performance attributes associated with URLLC services [29].

This is an extremely important parameter in 5G networks, corresponding to the time required for a data packet to be transmitted from the source to the destination, and essentially reflects the response time of the system.

Latency significantly influences the user experience and the feasibility of various applications and services. Its importance is particularly evident in meeting real-time demands imposed by critical applications.

In this context, the ability to ensure low-latency communication constitutes a fundamental pillar of 5G networks, enabling a wide range of new services and use cases with high reliability and performance requirements.

4.4.3. Reliability

Reliability is fundamental for URLLC services and is used to define QoS requirements [36], evaluate the performance of different multiplexing and resource allocation schemes [52], and compare algorithms [57].

This KPI is frequently expressed in the literature as the probability of successful and error-free data transmission [37,59] over a time interval or within a limited number of attempts. Essentially, it measures the consistency and assurance that the data sent will reach its destination within established QoS requirements.

Achieving the desired levels of reliability is crucial for the viability and success of many emerging use cases in 5G networks, particularly in critical scenarios in which data loss is not acceptable.

4.4.4. Fairness

Fairness, whose KPI is the Jain fairness index, is essential in resource management in 5G networks, seeking a balance between meeting the requirements of all types of traffic and maximizing spectral efficiency [8,24]. This ensures that no group of users is significantly disadvantaged in terms of data rate or other performance parameters [15,46]. Without this metric, critical services can monopolize network resources, severely compromising the performance of other services [19].

Thus, it is a highly relevant indicator in the context of the coexistence of different types of traffic, ensuring the equitable distribution of network resources, and maintaining acceptable service levels for all users or service classes. This prevents excessive degradation of the performance of certain flows in favor of others, promoting the balanced and efficient management of available resources.

4.4.5. Energy Efficiency

EE is a crucial KPI in the context of 5G mobile networks, particularly in IoT systems and wireless network optimization. This metric aims to provide a holistic view of the performance of solutions proposed in the literature, going beyond metrics focused exclusively on transmission rate or latency. EE validates whether the improvements achieved are sustainable from an energy perspective and whether the proposed solutions are feasible for implementation in real 5G networks [5,35,39].

Thus, approaches that demonstrate good energy efficiency while satisfying other QoS requirements offer a more robust and relevant validation of the strategies presented.

4.4.6. Block Error Rate

BLER is a KPI directly associated with the reliability of data transmission and is of utmost importance in 5G communication networks because it represents the proportion of data blocks transmitted with errors that require retransmission. Therefore, it is essential that BLER remains low to ensure excellent QoS, as it is a direct metric that quantifies reliability at the connection level. BLER is essential in critical applications such as industrial automation,

autonomous vehicles, and telemedicine, which require an extremely low probability of packet loss and can be used as an indicator of channel quality [15,52]. Although reliability is paramount for critical services, it also affects the transmission rates of all services [18,24,53].

Thus, BLER constitutes a crucial KPI in 5G networks, evaluating the reliability of the physical layer, and having a direct impact on the QoS experienced by users across various types of traffic. Supervision and regulation are vital for meeting reliability standards, maximizing resource utilization, and facilitating the effective integration of diverse traffic in 5G networks [35,48,76].

4.4.7. Spectral Efficiency

Spectral efficiency constitutes a crucial key performance indicator (KPI) in 5G networks, considering the constraints of the radio frequency spectrum and the need to accommodate the growing demand for data, along with a diverse range of services with distinct requirements.

Its importance is evidenced by various essential factors, including the optimization of spectrum utilization, increase in network capacity, facilitation of service coexistence, evaluation and comparison of resource allocation strategies, analysis of the impacts of network parameters, and weighing of gains and potential trade-offs against other KPIs [18,19,46,48,52].

Spectral efficiency is fundamental for evaluating the performance and sustainability of 5G networks, because it demonstrates the effectiveness with which limited spectral resources provide high-quality services to multiple users. Improving this KPI is essential to ensure the capacity, scalability, and economic viability of the networks. Consequently, research demonstrating high spectral efficiency combined with compliance with various QoS requirements provides greater robustness and relevance to the proposed solutions.

4.4.8. Effectiveness of the Techniques Presented in Section 4.3 in Relation to the KPIs

The effectiveness of the techniques presented in Section 4.3 in relation to the KPIs presented in this section is evaluated not only by their success in meeting the characteristics of conflicting KPIs of heterogeneous services but also by the trade-offs resulting from the underlying interaction. Table 4 summarizes this relationship. It is based on the evidence reported in the analyzed articles. The most impacted KPIs, the main trade-offs observed, and the reported limitations are discussed.

Table 4. Relationship between Scheduling/Allocation Techniques, Impacted KPIs, Trade-offs, and Key Limitations.

Methods	Most Impacted KPIs	Main Trade-Offs	Key Limitations
Puncturing	Latency and reliability (URLLC), throughput (eMBB)	Absolute priority to URLLC (low latency and high reliability) versus guaranted degradation of eMBB throughput.	High complexity of dynamic management; dependence on precise Channel State Information (CSI) for optimized decisions.
Superposition	Throughput, spectral efficiency, latency, reliability	It increases spectral efficiency and helps contain URLLC latency. However, it raises interference and decoding complexity. It can cause reliability degradation without proper power control.	Requires very precise power control and complex receivers (SIC); it is sensitive to CSI.

Table 4. *Cont.*

Methods	Most Impacted KPIs	Main Trade-Offs	Key Limitations
Heuristic Algorithm	Fairness, throughput, latency	Speed and low computational complexity versus renunciation of global optimality, resulting only in satisfactory solutions or local optima.	Difficulty in ensuring strict QoS requirements in high-load scenarios. Performance may degrade in very heterogeneous environments.
Network Slicing	Throughput, latency, reliability, spectral efficiency	Logical isolation and QoS and Service Level Agreement (SLA) guaranties for each service versus orchestration complexity and interference between slices; possible resource underutilization.	Complexity of management and orchestration of multiple slices; security challenges and interference between slices.
NOMA	Spectral efficiency, throughput, fairness	Fairness between users may be compromised; greater complexity in the receiver (requiring SIC)	Complexity in the implementation of the SIC; highly sensitive to the conditions of the CSI; difficulties in energy distribution.
Matching Theory	Fairness, latency, reliability	Optimization of equitable allocation under constraints vs. Overload and complexity of the model in large networks	High model complexity. Significant computational overload, especially in real-time. Strong dependence on CSI.
Iterative Algorithm	Energy Efficiency, throughput, latency, reliability	Obtaining viable (suboptimal) solutions for NP-hard problems versus guaranteeing convergence and stability.	Convergence and stability; strong sensitivity to initial parameters; scalability challenges in large-scale scenarios.
PPO/DRL	Throughput, latency, reliability, energy efficiency, fairness	Adaptive and multi-criteria optimization versus training and stability requirements.	Difficulty in training in complex environments; strong dependence on an adequate reward function; slow convergence and high computational overhead.

N.B.: This table summarizes the qualitative trends documented in the reviewed literature. The numerical values of the KPIs depend on the simulated scenarios, network parameters, and assumptions adopted in each study.

4.5. Impact and Future Trends

The discussion on future trends is based on the observations from Table 4, which synthesizes the available evidence on which KPIs each scheduling and allocation technique affects most significantly. It is observed that traditional techniques, such as Puncturing and overlay, are effective in optimizing specific KPIs but introduce significant trade-offs in other indicators. Non-orthogonal approaches, such as NOMA, increase spectral efficiency but require more complex power control and decoding mechanisms. Only AI-based techniques, such as DRL/PPO, demonstrate the ability to simultaneously optimize multiple conflicting KPIs (for example, eMBB throughput and URLLC latency), albeit at the cost of design and training complexity.

These observations justify the future trend, detailed below, of developing hybrid solutions that combine the computational efficiency of heuristics with the adaptability of

AI algorithms, as well as the adoption of architectures like Network Slicing, to provide fundamental QoS isolation.

4.5.1. Impact and Future Trends of Heterogeneous eMBB and URLLC Services Coexistence

The impact of service coexistence between eMBB and URLLC in 5G networks can be summarized as follows:

- Spectral resource scarcity, resulting from the need to meet their conflicting requirements [16,19];
- The interruption of eMBB transmissions caused by URLLC traffic, with an increase in URLLC users, significantly degrades the eMBB performance [16,19,59];
- The lack of precise channel information and dedicated bandwidth for URLLC poses a substantial barrier to meeting the strict latency requirements [17,27];
- Challenges in scheduling and optimal resource allocation, whose complexity increases with the need to ensure fairness and efficiency in resource sharing between services with contrasting QoS requirements [30,53,57].

Future trends in heterogeneous service coexistence (eMBB and URLLC) in 5G networks point to the development of innovative solutions based on the following three fundamental pillars:

1. AI/DRL algorithms for dynamic decision-making support.
2. Advanced multiplexing techniques and adaptive resource allocation.
3. Robust implementation of network slicing.

This multidimensional approach aims to achieve an optimal balance between the conflicting requirements of eMBB and URLLC services, enabling more efficient network resource management, and supporting a broader ecosystem of emerging applications.

4.5.2. Impact and Future Trends of Heterogeneous eMBB and mMTC Services Coexistence

The coexistence of eMBB and mMTC services in 5G networks highlights the need for:

- First Ensure a low Block Error Rate (BLER) for mMTC devices in short-packet communications [20];
- Manage congestion in Random Access (RA) scenarios, particularly when many mMTC devices simultaneously attempt to access the network while eMBB services are active [20,29,56];
- Address the complexity of resource allocation due to differing performance requirements [22,38,41,44];
- Mitigate the interference effects of mMTC on eMBB throughput, particularly when nonorthogonal access schemes are used [13,20,44].

In this context, the coexistence of heterogeneous services poses significant challenges in efficiently managing resource distribution and mitigating interference.

Future perspectives for eMBB and mMTC coexistence in 5G networks suggest the following:

- Network slicing to ensure logical isolation between services.
- NOMA to increase spectral efficiency.
- MIMO to optimize channel performance.
- AI/ML techniques, particularly DRL, are used for dynamic resource optimization to efficiently and adaptively meet the distinct requirements of both services, ensuring an appropriate QoS for each

4.5.3. Impact and Future Trends of Heterogeneous URLLC and mMTC Services Coexistence

The coexistence of URLLC and mMTC in 5G networks presents four fundamental challenges:

- Managing conflicting requirements with limited resources, such as mMTC prioritizes scalability and massive connectivity (up to 1 million devices/km^2), whereas URLLC demands latencies below 1 ms with 99.999% reliability [35,38,77];
- Interference at the physical layer stemming from mixed numerologies, mMTC uses a reduced subcarrier spacing (15/30 kHz) to maximize spectral efficiency, in contrast to the wider spacing (60/120 kHz) required by URLLC to minimize latency [5,56];
- Ensuring differentiated QoS requires dynamic allocation mechanisms that simultaneously ensure low-energy consumption for mMTC devices and ultrashort transmission windows for URLLC [29,35,77];
- Congestion in random access is particularly critical in massive access scenarios (10^6 devices/km^2), where resource contention can compromise the strict Service Level Agreements (SLAs) established for URLLC [54–56].

Thus, the coexistence of URLLC and mMTC faces significant challenges owing to the conflicting QoS requirements. Current research trends suggest adopting innovative approaches, such as:

- Network slicing enables the creation of independent virtual network segments dedicated to each type of service.
- NOMA, facilitating efficient multiple access and overcoming limitations of traditional orthogonal schemes.
- Critical mMTC, an emerging variant combining high reliability requirements with massive connectivity.
- Cooperative techniques, between base stations and terminal devices for distributed resource optimization.
- AI/ML techniques, aimed at dynamic and predictive management of network parameters.

4.5.4. Impact of Coexistence of eMBB, URLLC, and mMTC and Future Trends

The coexistence of eMBB, URLLC, and mMTC represents one of the primary goals of 5G and future 6G networks and is currently the focus of research. The main challenge lies in simultaneously satisfying distinct QoS requirements associated with each service. This combination integrates and reflects the challenges and future perspectives identified in the previously analyzed bilateral combinations.

Thus, the coexistence of these three services represents a multifaceted challenge that is currently being addressed through various sophisticated methodologies that emphasize adaptive and intelligent resource distribution, logical service segregation, and the exploration of new multiple-access modalities. Emerging trends indicate a transition toward increasingly dynamic, hybrid, flexible, and context-sensitive solutions driven by Artificial Intelligence and Machine Learning (AI/ML) and pioneering network structures such as network slicing and MEC.

4.6. Comparison of This Review Article on Service Coexistence in 5G Networks with a Previous Study in [5]

In contrast to previous reviews, particularly those conducted by the authors in [5], which focused exclusively on the simultaneity of service pairs (eMBB–URLLC), this study categorizes three additional types of mMTC. These findings reveal promising directions for future work, the evolution of coexistence mechanisms, and the key priorities for ongoing research. Although both articles are scoping reviews addressing the challenges of service

coexistence in 5G networks, they differ significantly in their scope, volume of literature analyzed, and specific emphases. Table 5 presents a detailed comparison of these two.

Table 5. Comparison of this review article on service coexistence in 5G networks with a previous.

Criteria		Previous Study [5]	This Work
Included articles		203	48
Period		2018–2022	2019–2024
Focus		The coexistence of eMBB and URLLC services within the 5G NR architecture, with a specific focus on 3GPP specifications and physical resource allocation methods.	Analyze and categorize the existing methods, approaches, and techniques for traffic scheduling and resource allocation among heterogeneous services (eMBB, URLLC, and mMTC) in 5G networks and beyond, with an emphasis on ensuring QoS and maximizing user satisfaction
Simultaneity of services	eMBB + URLLC	Yes	Yes
	eMBB + mMTC	No	Yes
	URLLC + mMTC	No	Yes
	eMBB + URLLC + mMTC	No	Yes
Discovered gaps		A more in-depth analysis of existing approaches for eMBB-URLLC coexistence, along with a detailed examination of the wide range of technical challenges in this context.	Explicit identification of a significant research gap regarding mMTC and its reliability, offering a more comprehensive view of the three 5G service types.
Final Summary		this study provides a comprehensive examination of current methodologies for eMBB-URLLC integration and enumerates various technical obstacles in this domain.	Research predominantly addresses eMBB and URLLC coexistence, while mMTC, vital for 6G, receives insufficient attention. Future methodologies may necessitate hybrid strategies incorporating AI/DRL, sophisticated multiplexing, NOMA, slicing, and integrated KPIs to navigate coexistence trade-offs.

Furthermore, explicit negligence regarding the reliability of mMTC was identified in the included studies. Although mMTC typically does not require high reliability, ignoring this requirement can severely compromise the continuous connectivity of critical sensors and actuators in Industry 4.0, smart cities, IoT, and future 6G applications. Inadequate concurrency between mMTC and critical services can lead to congestion, increased latency, and access failures, especially in dense environments [56,78,79]. The absence of reliability guarantees may result in data loss or unacceptable delays in industrial applications, automation, and environmental monitoring where real-time responses are essential [80].

It is essential to highlight that service heterogeneity is a core enabler of urban innovation in smart cities, as it allows for simultaneous responses to high-speed demands, low-latency requirements, and the connection of a massive number of devices. Investi-

gating this concurrency is crucial to ensuring service quality, optimizing resource allocation, and enabling critical and innovative applications in intelligent urban environments. Therefore, without service heterogeneity, the advancement of smart city applications is significantly hindered.

4.7. Importance and Impact of CSI on Scheduling and Resource Allocation Techniques

Accurate and appropriate information about CSI is essential for the efficient operation of scheduling and resource allocation techniques in 5G NR, particularly in heterogeneous service coexistence scenarios. The accuracy, frequency, and cost of obtaining CSI directly affect the efficiency, capacity, and energy consumption of the scheduling algorithms.

The procurement of CSI can manifest in a multitude of forms: instantaneous, statistical, predicted, and partial or obsolete. Each one of the exerts a distinct impact on scheduling. Instantaneous CSI provides maximum performance but with high overhead [81], while statistical CSI reduces overhead while maintaining performance close to ideal. Predicted CSI reduces overhead but relies heavily on the accuracy of the prediction [82], and partial or outdated CSI can significantly degrade performance, requiring adaptation mechanisms [83]

The practical implementation of the scheduling techniques reviewed in this study is highly dependent on the availability of accurate CSI. However, its acquisition in realistic smart city scenarios, characterized by high mobility, vehicular communications, and millimeter-wave (mmWave) bands, faces fundamental challenges that are often overlooked in the literature, which frequently assumes perfect and instantaneous CSI [21,24,27]

The primary CSI challenges relevant to scheduling are the following ones:

- Fast Channel Variation: High user mobility causes the channel to change rapidly, making accurate instantaneous CSI difficult to obtain and increasing the signaling overhead required to track it [84,85].
- Dynamic Urban Environments: Frequent signal blockages from buildings and obstacles in dense urban areas lead to unpredictable channel variations, complicating modeling and prediction [84,86].

The underlying impact of CSI on the applied scheduling techniques are as follows:

- Puncturing and Superposition: Dynamic techniques that rely on precise channel knowledge to decide where to puncture or how to overlay signals without causing catastrophic interference. In vehicular environments, CSI can become obsolete between measurement and transmission, resulting in incorrect decisions for puncturing or overlaying, which degrade eMBB performance and generate more interference than benefit.
- Network Slicing: The reliability of CSI is crucial for resource forecasting and ensuring isolation at the slice level. Imprecise CSI can lead to resource over-selling and SLA violations, especially in critical services like URLLC.
- DRL/PPO: These approaches can offer greater robustness in the face of imperfect CSI, as they learn policies based on partial or historical states of the channel. However, the performance during the training phase can be biased by incorrect estimates, resulting in suboptimal decisions during operation in dynamic traffic scenarios [27,37].
- NOMA and Matching Theory: They are highly sensitive to the quality of the CSI. In NOMA, inaccurate CSI compromises the efficiency of SIC and increases error propagation. In Matching Theory, the calculation of utilities for stable matching requires precise channel measurements. Outdated CSI can lead to severe suboptimization and fairness breakdown [21,24].

As demonstrated, the reliability of CSI is a critical factor that can enhance or limit the effectiveness of any scheduling technique. Therefore, its choice for deployment in realistic smart city scenarios must critically consider its robustness to CSI uncertainty. Future

research directions should explicitly explore the integration of robust channel estimation models (such as those based on deep learning) with scheduling algorithms and evaluate the performance of these techniques under conditions of imperfect and delayed CSI, and not just under ideal conditions.

4.8. Limitations of the Review

This review presents several limitations. First, the search was restricted to two databases (IEEE Xplore and Scopus), which, although highly relevant and specialized for the domains of 5G networks (including smart cities) and telecommunications, may have limited the comprehensiveness of the literature analyzed. Second, the review protocol was not pre-registered on public platforms such as OSF or INPLASY, potentially affecting the transparency and traceability of the methodology. Third, data extraction was conducted by a single reviewer, which may introduce bias in the selection or synthesis of information. Fourth, no critical appraisal of the methodological quality of the included studies was carried out, as this is not a mandatory requirement for scoping reviews. Lastly, only studies published in English were included, which may have resulted in the exclusion of relevant research published in other languages, thereby introducing potential language bias.

Additionally, our analysis indicated that the acquisition of CSI in high-mobility and mmWave contexts constitutes a substantial limitation on scheduling performance, despite not being part of our initial search parameters. We chose our keywords based on basic scheduling methods (scheduling, coexistence, eMBB, URLLC, mMTC, 5G) rather than specific channel conditions. Consequently, our review may not have sufficiently encompassed scheduling studies that focus on mobility constraints and sophisticated channel estimation techniques, suggesting a fruitful avenue for future investigation.

5. Final Considerations

This scoping review highlights that existing scientific literature predominantly emphasizes the coexistence of eMBB and URLLC services, whereas mMTC, integrated with other service paradigms, remains insufficiently explored. This gap is particularly evident in critical scenarios such as industrial networks, smart cities, and digital health, which are poised to be fundamental in the 6G ecosystem.

These results suggest that future coexistence mechanisms will require greater complexity and sophistication. They are expected to evolve into hybrid methodologies, integrating AI/DRL algorithms with reduced computational demands, advanced multiplexing techniques, dynamic resource allocation, network slicing, and NOMA for effective interference management, along with energy-efficient solutions capable of adapting to imperfect channel conditions.

Furthermore, key priorities for future research include exploring triple coexistence (eMBB + URLLC + mMTC), particularly in industrial contexts, using hybrid methodologies (e.g., DRL combined with network slicing). In addition, the establishment of integrated KPIs that encapsulate the trade-offs between scalability, connectivity, throughput, latency, and energy efficiency is essential. Ultimately, the formulation of flexible architectures capable of reconciling conflicting requirements without compromising the QoS is imperative.

In conclusion, the transition to 6G will require innovative solutions that ensure the efficient coexistence of heterogeneous services and integrate artificial intelligence, resource optimization, and energy sustainability. This study underscores the need for ongoing research in this domain, especially in contexts that are still underexplored but critical for the future of communication networks.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/smartcities8050168/s1, File S1.

Author Contributions: Conceptualization, N.R.P. and F.J.V.; methodology, N.R.P. and F.J.V.; software, N.R.P.; validation, X N.R.P. and F.J.V.; formal analysis, N.R.P. and F.J.V.; investigation N.R.P. and F.J.V.; resources, N.R.P.; data curation, N.R.P.; writing—original draft preparation, N.R.P.; writing—review and editing, F.J.V.; visualization, N.R.P.; supervision, F.J.V.; project administration, F.J.V.; funding acquisition, F.J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Fundação para a Ciência e a Tecnologia, I.P. (FCT—Portuguese Foundation for Science and Technology), through national funds (FCT/MECI) and, when applicable, co-funded by EU funds under project UID/50008—Instituto de Telecomunicações.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Acknowledgments: Ntunitangua also thanks INAGBE for the support provided through the doctoral scholarship for faculty staff, awarded via the Instituto Superior Politécnico de Ndalatando (ISPNd). This research was also supported by COST Action CA20120 INTERACT—Intelligence-Enabling Radio Communications for Seamless Inclusive Interactions. The authors would also like to acknowledge the fruitful discussions with Ivan Miguel Pires from ESTGA, University of Aveiro, and Instituto de Telecomunicações.

Conflicts of Interest: The authors declare that they have no conflicts of interest related to this study.

Abbreviations

The following abbreviations are used in this manuscript:

5G	Fifth Generation
5G NR	Fifth Generation New Radio
6G	Sixth Generation
AI	Artificial Intelligence
B5G	Beyond 5G
BCD	Block Coordinate Descent
BLER	Block Error Rate
CN	Core Network
CSI	Channel State Information
CTMC	Continuous Time Markov Chain
CVaR	Conditional Value-at-Risk
D2D	Device-to-Device
DDGSP	Data-Driven Genetic Algorithm-Based Spectrum Partition
DDQN	Deep Double Q-Learning
DQN	Deep Q-Network
DRAPS	Dynamic Resource Allocation and Puncturing Strategy
DRL	Deep Reinforcement Learning
DROA	Decomposition-Relaxation-Optimization Algorithm
E2E	End-to-End
EE	Energy Efficiency
eMBB	enhanced Mobile Broadband
GNNs	Graph Neural Networks
GS	Gale-Shapley
H-CRAN	Heterogeneous Cloud Radio Access Networks
HMA	Hybrid orthogonal/non-orthogonal Multiple Access
H-OMA/H-NOMA	Hybrid OMA/Hybrid NOMA
INPLASY	International Platform of Registered Systematic Review and Meta-analysis Protocols
IoT	Internet of Things
ITU	International Telecommunication Union
JRCRA	Joint Radio and Core Resource Allocation

LRT-Q	Latency-Reliability-Throughput Improvement in 5G NR using Q-Learning
MAC	Medium Access Control
MBS	Macro Base Station
MC	Multi-connectivity/Markov Chain
MEAR	Minimum Expected Achieved Rate
MEC	Multi-access Edge Computing
MIMO	Multiple-Input Multiple-Output
MINLP	Mixed Integer Nonlinear Programming
ML	Machine Learning
mMTC	massive Machine Type Communication
MNO	Mobile Network Operator
MOO	Multi-Objective Optimization
mRBs	Mini Resource Blocks
MVNOs	Mobile Virtual Network Operators
NF-FN	Near-Far/Far-Near
NFV	Network Function Virtualization
NN-FF	Near-Near/Far-Far
NOMA	Non-Orthogonal Multiple Access
OMA	Orthogonal Multiple Access
OPM	Objective Product Method
OSF	Open Science Framework
OSSPA	Optimized Sparrow Search Algorithm
PI	Preemption Indication
PLR	Packet Loss Ratio
PPF	Personalized Performance Fluctuation
PPO	Proximal Policy Optimization
PRB	Physical Resource Block
PRISMA-SCR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PSO	Particle Swarm Optimization
QoE	Quality of Experience
QoS	Quality of Service
RACH	Random Access Channel
RAN	Radio Access Network
RBs	Resource Blocks
RIS	Research Information Systems
RSMA	Rate-Splitting Multiple Access
SAFE-TS	Self-adaptive Flexible Transmission Time Interval Scheduling
SBS	Small Base Station
SCA	Successive Convex Approximation
SE	Spectral Efficiency
SIC	Successive Interference Cancellation
SINR	Signal-to-Interference-plus-Noise Ratio
SJF	Shortest Job First
SLA	Service Level Agreements
SOO	Single-Objective Optimization
SSR	Service Level Agreement Satisfaction Ratio
TD3	Twin Delayed Deep Deterministic Policy Gradient
TS	Traffic Steering
TSBART	Task Scheduling, Bandwidth Allocation, and Robot Trajectory
TTI	Transmission Time Interval
URLLC	Ultra-reliable low-latency communication

References

1. ITU-R M.2083-0, "IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond". 2015. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf (accessed on 11 June 2024).
2. Almekhlafi, M.; Arfaoui, M.A.; Assi, C.; Ghrayeb, A. Superposition-Based URLLC Traffic Scheduling in 5G and Beyond Wireless Networks. *IEEE Trans. Commun.* **2022**, *70*, 6295–6309. [CrossRef]
3. Haque, E.; Tariq, F.; Khandaker, M.R.A.; Wong, K.-K.; Zhang, Y. A Survey of Scheduling in 5G URLLC and Outlook for Emerging 6G Systems. *IEEE Access* **2023**, *11*, 34372–34396. [CrossRef]
4. Yin, H.; Zhang, L.; Roy, S. Multiplexing URLLC Traffic within eMBB Services in 5G NR: Fair Scheduling. *IEEE Trans. Commun.* **2020**, *69*, 1080–1093. [CrossRef]
5. Kumar, R.; Sinwar, D.; Singh, V. QoS aware resource allocation for coexistence mechanisms between eMBB and URLLC: Issues, challenges, and future directions in 5G. *Comput. Commun.* **2023**, *213*, 208–235. [CrossRef]
6. Paz-Pérez, A.; Tato, A.; Escudero-Garzás, J.J.; Gómez-Cuba, F. Flexible Reinforcement Learning Scheduler for 5G Networks. In Proceedings of the 2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN), Stockholm, Sweden, 5–8 May 2024; pp. 566–572.
7. Pamba, R.V.; Bhandari, R.; Asha, A.; Neware, R.; Bist, A.S. Novel Deep Learning Approach to Support Optimal Resource Allocation in 5G Environment. *J. Mob. Multimed.* **2023**, *19*, 739–763. [CrossRef]
8. Bischoff, T.; Kasparick, M.; Tohidi, E.; Stańczak, S. Real-Time Algorithms for Combined eMBB and URLLC Scheduling. In Proceedings of the 2024 27th International Workshop on Smart Antennas (WSA), Dresden, Germany, 17–19 March 2024; pp. 1–5.
9. Wang, L.; Tao, S.; Zhao, L.; Zhou, D.; Liu, Z.; Sun, Y. Resource Scheduling in URLLC and eMBB Coexistence Based on Dynamic Selection Numerology. *Wirel. Commun. Mob. Comput.* **2024**, *2024*, 9480388. [CrossRef]
10. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
11. A Akl, E.; Khabsa, J.; Iannizzi, C.; Piechotta, V.; A Kahale, L.; Barker, J.M.; E McKenzie, J.; Page, M.J.; Skoetz, N. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): Checklist and explanation. *BMJ* **2024**, *387*, e079183. [CrossRef] [PubMed]
12. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef]
13. Nadif, S.; Sabir, E.; Elbiaze, H.; Habachi, O.; Haqiq, A. A Hierarchical Green Mean-Field Power Control with eMBB-mMTC Coexistence in Ultradense 5G (Invited Paper. In Proceedings of the 2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt), Turin, Italy, 19–22 September 2022; pp. 330–337.
14. Zhang, J.; Li, L. A Greedy Strategy of Multiplexing uRLLC Traffic Within eMBB Services for HSR. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; pp. 1670–1674.
15. Elsayed, M.; Erol-Kantarci, M. AI-Enabled Radio Resource Allocation in 5G for URLLC and eMBB Users. In Proceedings of the 2019 IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, 30 September–2 October 2019; pp. 590–595.
16. Bairagi, A.K.; Munir, S.; Alsenwi, M.; Tran, N.H.; Hong, C.S. A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks. In Proceedings of the SAC '19: The 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; pp. 2377–2384.
17. Zhu, X.; Wang, J.; Li, J.; Lu, H.; Luo, X.; Lai, Q. An Approach to Transmitting URLLC Data with Different Latency Requirements over eMBB Services Based on Deep Reinforcement Learning. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 10–13 December 2021; pp. 120–124.
18. Tian, M.; Li, C.; Hui, Y.; Chen, B.; Yue, W.; Fu, Y.; Han, Z. An Intelligent Coexistence Strategy for eMBB/URLLC Traffic in Multi-UAV Relay Networks via Deep Reinforcement Learning. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 13424–13439. [CrossRef]
19. Zhao, S.; Wang, Y.; Wang, T.; Wang, Z. A Reservation-Based Hybrid Multiple Access Scheme for URLLC Coexisting with eMBB. In Proceedings of the 2021 IEEE/CIC International Conference on Communications in China (ICCC), Xiamen, China, 28–30 July 2021; pp. 916–921.
20. Zhu, X.; Wang, J.; Li, J.; Lu, H.; Lai, Q.; Luo, X. A Scheme for Uplink NOMA Communication with Intelligent Resource Allocation for mMTC Traffic over eMBB Traffic. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring), Helsinki, Finland, 19–22 June 2022; pp. 1–5.
21. Chen, Q.; Wu, J.; Wang, J.; Jiang, H. Coexistence of URLLC and eMBB Services in MIMO-NOMA Systems. *IEEE Trans. Veh. Technol.* **2022**, *72*, 839–851. [CrossRef]
22. Hou, W.; Zhu, X.; Cao, J.; Zeng, H.; Jiang, Y. Composite Robot Aided Coexistence of eMBB, URLLC and mMTC in Smart Factory. In Proceedings of the 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall), London, UK, 26–29 September 2022; pp. 1–6.

23. Manzoor, A.; Kazmi, S.M.A.; Pandey, S.R.; Hong, C.S. Contract-Based Scheduling of URLLC Packets in Incumbent EMBB Traffic. *IEEE Access* **2020**, *8*, 167516–167526. [CrossRef]
24. Prathyusha, Y.; Sheu, T.-L. Coordinated Resource Allocations for eMBB and URLLC in 5G Communication Networks. *IEEE Trans. Veh. Technol.* **2022**, *71*, 8717–8728. [CrossRef]
25. Peng, H.; Wang, L.-C.; Jian, Z. Data-Driven Spectrum Partition for Multiplexing URLLC and eMBB. *IEEE Trans. Cogn. Commun. Netw.* **2022**, *9*, 386–397. [CrossRef]
26. Wenqi, Z.; Zhiwen, P.; Nan, L.; Xiaohu, Y. Deep Reinforcement Learning Based Dynamic Resource Slicing for eMBB and URLLC Traffic Considering Puncturing. In Proceedings of the 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), Singapore, 24–27 June 2024; pp. 1–6.
27. Saggese, F.; Pasqualini, L.; Moretti, M.; Abrardo, A. Deep Reinforcement Learning for URLLC data management on top of scheduled eMBB traffic. In Proceedings of the GLOBECOM 2021—2021 IEEE Global Communications Conference, Madrid, Spain, 7–11 December 2021; pp. 1–6.
28. Zhang, K.; Xu, X.; Zhang, J.; Zhang, B.; Tao, X.; Zhang, Y. Dynamic Multiconnectivity Based Joint Scheduling of eMBB and uRLLC in 5G Networks. *IEEE Syst. J.* **2020**, *15*, 1333–1343. [CrossRef]
29. Awada, Z.; El Helou, M.; Khawam, K.; Lahoud, S. Dynamic Multi-Tenant RAN Slicing for eMBB, URLLC, and mMTC in 5G Networks. In Proceedings of the 2023 26th International Symposium on Wireless Personal Multimedia Communications (WPMC), Tampa, FL, USA, 19–22 November 2023; pp. 1–7.
30. Wang, L.; Yuan, J.; Jiang, X.; Cui, J.; Zheng, B. Dynamic Resource Scheduling Strategy with QoE-aware for the Coexistence of eMBB and URLLC Traffic. In Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Virtual Meeting, China, 20–22 October 2021; pp. 1–5.
31. Yang, W.; Li, C.-P.; Fakoorian, A.; Hosseini, K.; Chen, W. Dynamic URLLC and eMBB Multiplexing Design in 5G New Radio. In Proceedings of the 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), La Vegas, USA, 10–13 January 2020; pp. 1–5.
32. Li, C.; Liu, B.; Su, X.; Xu, X. eMBB-URLLC Multiplexing: A Greedy Scheduling Strategy for URLLC Traffic with Multiple Delay Requirements. In Proceedings of the TENCON 2023–2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand, 31 October–3 November 2023; pp. 1216–1221.
33. Souza, C.; Falcão, M.; Balieiro, A.; Taleb, T.; Alves, E. Enabling the eMBB and URLLC coexistence in MEC-NFV Networks. In Proceedings of the ICC 2024—IEEE International Conference on Communications, Denver, CO, USA, 9–13 June 2024; pp. 159–164.
34. Taskou, S.K.; Rasti, M.; Hossain, E. End-to-End Resource Slicing for Coexistence of eMBB and URLLC Services in 5G-Advanced/6G Networks. *IEEE Trans. Mob. Comput.* **2023**, *23*, 8015–8032. [CrossRef]
35. Elgarhy, O.; Reggiani, L.; Alam, M.M.; Zoha, A.; Ahmad, R.; Kuusik, A. Energy Efficiency and Latency Optimization for IoT URLLC and mMTC Use Cases. *IEEE Access* **2024**, *12*, 23132–23148. [CrossRef]
36. Ren, R.; Wang, J.; Yu, J.; Zhu, X.; Wan, X.; Lu, H. Hybrid Puncturing and Superposition Scheme for Multiplexing uRLLC and eMBB Services Based on Deep Reinforcement Learning. In Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 9–12 December 2022; pp. 806–810.
37. Sohaib, R.M.; Onireti, O.; Sambo, Y.; Swash, R.; Ansari, S.; Imran, M.A. Intelligent Resource Management for eMBB and URLLC in 5G and Beyond Wireless Networks. *IEEE Access* **2023**, *11*, 65205–65221. [CrossRef]
38. Ren, R.; Wang, J.; Yu, J.; Zhu, X.; Wan, X.; Lu, H. Joint Resource Allocation for Multiplexing eMBB, URLLC and mMTC Traffics Based on DRL. In Proceedings of the 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), Singapore, 24–27 June 2024; pp. 1–5.
39. Cheng, Q.; Li, K.; Zhu, P.; Li, J.; Jiang, Y.; Wang, D. Joint Resource Block and Power Allocation for eMBB and URLLC Coexistence in 5G H-CRAN. In Proceedings of the 2023 International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, 2–4 November 2023; pp. 960–965.
40. Sun, H.; Yang, J.; Su, J.; Wang, H.; Liu, D. Joint Resource Scheduling for Coexistence of URLLC and eMBB in 5G Wireless Networks. In Proceedings of the 2021 Computing, Communications and IoT Applications (ComComAp), Shenzhen, China, 26–28 November 2021; pp. 53–58.
41. Almekhlafi, M.; Arfaoui, M.A.; Assi, C.; Ghrayeb, A. Joint Resource and Power Allocation for URLLC-eMBB Traffics Multiplexing in 6G Wireless Networks. In Proceedings of the ICC 2021—IEEE International Conference on Communications, Virtual/Montreal, Canada, 14–23 June 2021; pp. 1–6.
42. Liu, J.; Liu, B.; Su, X.; Yu, X.; Xu, X. Joint Scheduling Scheme for eMBB/URLLC Based on Multi-User Superposition Transmission. In Proceedings of the ICC 2024—IEEE International Conference on Communications, Denver, CO, USA, 9–13 June 2024; pp. 999–1004.
43. Zhang, J.; Xu, X.; Zhang, K.; Zhang, B.; Tao, X.; Zhang, P. Machine Learning Based Flexible Transmission Time Interval Scheduling for eMBB and uRLLC Coexistence Scenario. *IEEE Access* **2019**, *7*, 65811–65820. [CrossRef]

44. Tominaga, E.N.; Alves, H.; Lopez, O.L.A.; Souza, R.D.; Rebelatto, J.L.; Latva-Aho, M. Network Slicing for eMBB and mMTC with NOMA and Space Diversity Reception. In Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Online, Finland, 25 April–19 May 2021; pp. 1–6.
45. Liu, Y.; Clerckx, B.; Popovski, P. Network Slicing for eMBB, URLLC, and mMTC: An Uplink Rate-Splitting Multiple Access Approach. *IEEE Trans. Wirel. Commun.* **2023**, *23*, 2140–2152. [CrossRef]
46. Sekhar, R.C.; Singh, P. Optimization of resource allocation in 5G networks: A network slicing approach with hybrid NOMA for enhanced uRLLC and eMBB coexistence. *Int. J. Commun. Syst.* **2024**, *37*, e5928. [CrossRef]
47. Tian, M.; Li, C.; Hui, Y.; Cheng, N.; Luo, M. Optimized Sparrow Search-based Multiplexing of eMBB and URLLC in 5G/B5G Networks. In Proceedings of the GLOBECOM 2022—2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; pp. 3658–3663.
48. Ivanova, D.; Zhbankova, E.; Markova, E.; Gaidamaka, Y.; Samouylov, K. Performance modeling and comparison of URLLC and eMBB coexistence strategies in 5G new radio systems. *Comput. Netw.* **2024**, *255*, 110904. [CrossRef]
49. Guo, J.; Nie, G.; Tian, H.; Zhang, B.; Yuan, J. Puncture-Predictive Fairness Scheduling Scheme for eMBB and URLLC Based on TD3 Algorithm. In Proceedings of the 2023 IEEE/CIC International Conference on Communications in China (ICCC), Dalian, China, 10–12 August 2023; pp. 1–6.
50. Shi, B.; She, C.; Zheng, F.-C.; Gao, L.; Li, G. Puncturing-Based Resource Allocation for URLLC and eMBB Services via Matching Theory and Unsupervised Deep Learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 13396–13411. [CrossRef]
51. Dos Santos, E.J.; Souza, R.D.; Rebelatto, J.L. Rate-Splitting Multiple Access for URLLC Uplink in Physical Layer Network Slicing With eMBB. *IEEE Access* **2021**, *9*, 163178–163187. [CrossRef]
52. Zhao, Y.; Chi, X.; Qian, L.; Zhu, Y.; Hou, F. Resource Allocation and Slicing Puncture in Cellular Networks With eMBB and URLLC Terminals Coexistence. *IEEE Internet Things J.* **2022**, *9*, 18431–18444. [CrossRef]
53. Wang, P.; Jiang, C.; Mao, Z.; Chen, Y.; Liu, F. Resource allocation scheme to reduce computing energy consumption of uRLLC and eMBB services in MEC scenarios. In Proceedings of the 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 24–26 May 2024; pp. 1584–1589.
54. Shi, B.; Zheng, F.-C.; She, C.; Luo, J.; Burr, A.G. Risk-Resistant Resource Allocation for eMBB and URLLC Coexistence Under M/G/1 Queueing Model. *IEEE Trans. Veh. Technol.* **2022**, *71*, 6279–6290. [CrossRef]
55. Abreu, R.; Jacobsen, T.; Pedersen, K.; Berardinelli, G.; Mogensen, P. System Level Analysis of eMBB and Grant-Free URLLC Multiplexing in Uplink. In Proceedings of the 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 28 April–1 May 2019; pp. 1–5.
56. Pokhrel, S.R.; Ding, J.; Park, J.; Park, O.-S.; Choi, J. Towards Enabling Critical mMTC: A Review of URLLC Within mMTC. *IEEE Access* **2020**, *8*, 131796–131813. [CrossRef]
57. Daneshvar, S.M.M.H.; Mazinani, S.M. Training a Graph Neural Network to solve URLLC and eMBB coexisting in 5G networks. *Comput. Commun.* **2024**, *225*, 171–184. [CrossRef]
58. Zaki-Hindi, A.; Elayoubi, S.-E.; Chahed, T. URLLC and eMBB coexistence in unlicensed spectrum: A preemptive approach. In Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15–19 June 2020; pp. 229–234.
59. Almekhlafi, M.; Chraiti, M.; Arfaoui, M.A.; Assi, C.; Ghrayeb, A.; Alloum, A.; Hamood, M. A Downlink Puncturing Scheme for Simultaneous Transmission of URLLC and eMBB Traffic by Exploiting Data Similarity. *IEEE Trans. Veh. Technol.* **2021**, *70*, 13087–13100. [CrossRef]
60. Bairagi, A.K.; Munir, S.; Alsenwi, M.; Tran, N.H.; Alshamrani, S.S.; Masud, M.; Han, Z.; Hong, C.S. Coexistence Mechanism Between eMBB and uRLLC in 5G Wireless Networks. *IEEE Trans. Commun.* **2020**, *69*, 1736–1749. [CrossRef]
61. Liu, Y.; Guo, K.; Wang, Q. Fairness-Centric Resource Allocation Methods for eMBB and URLLC Services in 5G and Beyond Networks. In Proceedings of the GLOBECOM 2024—2024 IEEE Global Communications Conference, Cape Town, South Africa, 8–12 December 2024; pp. 3297–3302.
62. Lu, K.; Jiang, C. Optimized Low Density Superposition Modulation for 5G Mobile Multimedia Wireless Networks. *IEEE Access* **2019**, *7*, 174227–174235. [CrossRef]
63. Yeom, J.S.; Chu, E.; Jung, B.C.; Jin, H. Performance Analysis of Diversity-Controlled Multi-User Superposition Transmission for 5G Wireless Networks. *Sensors* **2018**, *18*, 536. [CrossRef]
64. Mastan, S.; Balakrishna, U.; Raju, G.S.S. Heuristic Algorithm Strategies to Solve Travelling salesman Problem. *J. Xi'an Univ. Archit. Technol.* **2020**, *7*, 2073–2079.
65. Al-Shaery, A.M.; Khozium, M.O.; Farooqi, N.S.; Alshehri, S.S.; Al-Kawa, M.A.M. Problem Solving in Crowd Management Using Heuristic Approach. *IEEE Access* **2022**, *10*, 25422–25434. [CrossRef]
66. Zhang, S. An Overview of Network Slicing for 5G. *IEEE Wirel. Commun.* **2019**, *26*, 111–117. [CrossRef]
67. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [CrossRef]

68. Guo, W.; Qureshi, N.M.F.; Siddiqui, I.F.; Shin, D.R. Cooperative Communication Resource Allocation Strategies for 5G and Beyond Networks: A Review of Architecture, Challenges and Opportunities. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 8054–8078. [CrossRef]
69. Shin, W.; Vaezi, M.; Lee, B.; Love, D.J.; Lee, J.; Poor, H.V. Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges. *IEEE Commun. Mag.* **2017**, *55*, 176–183. [CrossRef]
70. Ali, S.; Tabassum, H.; Hossain, E. Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems. *IEEE Access* **2016**, *4*, 6325–6343. [CrossRef]
71. Zhang, C.; Wu, C.; Lin, M.; Lin, Y.; Liu, W. Proximal Policy Optimization for Efficient D2D-Assisted Computation Offloading and Resource Allocation in Multi-Access Edge Computing. *Futur. Internet* **2024**, *16*, 19. [CrossRef]
72. Gatti, R.; G.B., A.K.; K.N., S.K.; Palle, S.; Gadekallu, T.R. Optimal resource scheduling algorithm for cell boundaries users in heterogenous 5G networks. *Phys. Commun.* **2022**, *55*, 101915. [CrossRef]
73. Alsenwi, M.; Tran, N.H.; Bennis, M.; Pandey, S.R.; Bairagi, A.K.; Hong, C.S. Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4585–4600. [CrossRef]
74. Umesh, R.G.; Sushil Kumar, G.N.; Santhosh, K.; Suraksha, M.S.; Praveen Kumar, K.V. Radio Resource Allocation for 5G Network Using Deep Reinforcement Learning. *Int. J. Res. Appl. Sci. Eng. Technol.* **2023**, *11*, 677–683. [CrossRef]
75. Al-Tam, F.; Correia, N.; Rodriguez, J. Learn to Schedule (LEASCH): A Deep Reinforcement Learning Approach for Radio Resource Scheduling in the 5G MAC Layer. *IEEE Access* **2020**, *8*, 108088–108101. [CrossRef]
76. Yang, K.; Zeng, Y.; Jiang, H.; Chen, Q. Cognitive Hierarchy Based Coexistence and Resource Allocation for URLLC and Embb. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2019; Volume 11910 LNCS, pp. 150–160. [CrossRef]
77. Sabuj, S.R.; Ahmed, A.; Cho, Y.; Lee, K.-J.; Jo, H.-S. Cognitive UAV-Aided URLLC and mMTC Services: Analyzing Energy Efficiency and Latency. *IEEE Access* **2021**, *9*, 5011–5027. [CrossRef]
78. Kalor, A.E.; Durisi, G.; Coleri, S.; Parkvall, S.; Yu, W.; Mueller, A.; Popovski, P. Wireless 6G Connectivity for Massive Number of Devices and Critical Services. *Proc. IEEE* **2024**, 1–23. [CrossRef]
79. Liu, L.; Larsson, E.G.; Popovski, P.; Caire, G.; Chen, X.; Khosravirad, S.R. Guest Editorial: Massive Machine-Type Communications for IoT. *IEEE Wirel. Commun.* **2021**, *28*, 56. [CrossRef]
80. Kovtun, V.; Kovtun, O.; Grochla, K.; Yasniy, O. The quality of service assessment of eMBB and mMTC traffic in a clustered 5G ecosystem of a smart factory. *Egypt. Inform. J.* **2025**, *29*, 100598. [CrossRef]
81. Gu, Z.; Hardjawana, W.; Vucetic, B. Opportunistic Scheduling Using Statistical Information of Wireless Channels. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 9810–9825. [CrossRef]
82. Wei, J.; Ye, D. Multisensor Scheduling for Remote State Estimation Over a Temporally Correlated Channel. *IEEE Trans. Ind. Inform.* **2022**, *19*, 800–808. [CrossRef]
83. Zeng, F.; Zhang, R.; Cheng, X.; Yang, L. Channel Prediction Based Scheduling for Data Dissemination in VANETs. *IEEE Commun. Lett.* **2017**, *21*, 1409–1412. [CrossRef]
84. Chen, Y.; Wang, Y.; Jiao, L. Robust Transmission for Reconfigurable Intelligent Surface Aided Millimeter Wave Vehicular Communications With Statistical CSI. *IEEE Trans. Wirel. Commun.* **2021**, *21*, 928–944. [CrossRef]
85. You, C.; Zhang, R. Hybrid Offline-Online Design for UAV-Enabled Data Harvesting in Probabilistic LoS Channels. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3753–3768. [CrossRef]
86. Zivuku, P.; Adam, A.B.M.; Ntontin, K.; Kisseleff, S.; Ha, V.N.; Chatzinotas, S.; Ottersten, B. Geographical Fairness in Multi-RIS-Assisted Networks in Smart Cities: A Robust Design. *IEEE Trans. Commun.* **2025**, *73*, 6622–6638. [CrossRef]