This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



# Estimating Latency-Reliability in B5G Radio Access Networks: an AI-Empowered Approach

Francesca Conserva*, Carolina Gijón †, Matías Toril†, Davide Micheli‡, Maurizio Fodrini§, and Roberto Verdone*
* Department of Electrical, Electronics, and Information Engineering (DEI), University of Bologna & WiLab (CNIT), *Italy*
† Telecommunication Research Institute (TELMA), University of Málaga, *Spain*
‡ TIM, *Italy*
§ FiberCop, *Italy*

***Abstract*—The challenge of ultra-low latency is boosting the interest of Mobile Network Operators (MNOs) in advanced mitigation strategies relying on latency decomposition to trace root causes. This work proposes the first hybrid framework combining analytical and data-driven modeling to estimate Latency-Reliability (LR), defined as the percentage of packets received within a target latency threshold, for different service classes in cellular networks. A preliminary full-stack analysis of Next Generation NodeB operation, coupled with a queuing-theoretic latency model, identifies critical delay stages and maps them to relevant Key Performance Indicators (KPIs) from the wide range of network measurements collected in the Operational Support System. These KPIs feed a module that integrates Supervised Learning, expert knowledge, and statistical analysis to predict LR per service class at cell level, empowering MNOs to take proactive actions in latency-sensitive scenarios. Assessment conducted over real network data shows that the proposed approach achieves latency estimation errors below 10% across service classes, allowing for accurate LR predictions for voice traffic in the absence of ground-truth data. Additionally, dimensionality reduction improves both computational efficiency and model interpretability, supporting targeted latency mitigation.**

***Index Terms*—Cellular Networks, 5G, Latency, Reliability, Estimation, KPI, Supervised Learning.**

## I. Introduction

THE 5th generation of mobile communications is reshaping technology and society, expanding Mobile Network Operators (MNOs)'s business from end-users to vertical industries [1]. 5th Generation (5G) systems combine the use of new frequency bands with features such as Network Function Virtualization (NFV), Software Defined Networkings (SDNs) [2], Multi-Edge Computing (MEC) [3] and Network Slicing (NS) to offer peak data rates of up to 10 Gbps, massive device connectivity and End-to-End (E2E) latency below 1 ms. Yet, 5G capabilities are increasingly strained by next-generation applications. Innovative services such as digital twins, holographic telepresence, extended reality and tele-surgery are driving breakthroughs across sectors such as healthcare, entertainment and smart cities, fostering a seamless convergence of physical and digital worlds [4]. These advancements demand a redefinition of 5G use cases, with Ultra-Reliable Low-Latency Communications (URLLC) evolving into Immersive Hyper-Reliable Low Latency Communication (IHRLLC), paving the way for 6th Generation (6G) and its near-instantaneous connectivity [5]. As a result, the challenge of ultra-low latency is inherited and intensified in 6G, evincing the need for refined mitigation strategies aware of key latency components [6].

Decomposing Radio Access Network (RAN) latency remains particularly difficult, posing a critical challenge for MNOs. This complexity stems from the heterogeneous network topologies (i.e., macro vs. small cells, aerial vs. terrestrial base station), the wide range of service offered to users with very different handset capabilities, the varying service mix across cells, and vendor-specific solutions for Radio Resource Management (RRM) procedures such as packet scheduling and slice resource allocation, whose details are often unknown by MNOs. Additional factors such as interference, spectrum availability, fluctuating radio link quality and mobility make RAN latency modeling far more intricate than in more stable and centralized domains like the core network [7], [8], [9]. Despite these difficulties, isolating latency critical stages within the RAN remains crucial for Service Level Agreement (SLA) fulfillment in Beyond 5G (B5G) networks.

Some prior works have attempted to address RAN latency decomposition. The existing literature predominantly focuses on modeling the User (U)-Plane latency [10] in the RAN. Contributions such as [11], [12], [13] provide mathematical breakdowns of delay components, identifying sources of delay within protocol stacks and transmission paths. However, they fall short in offering MNOs actionable insights that allow targeted interventions. To date, the literature lacks analytical models linking RAN latency components to measurable cell-level Key Performance Indicators (KPIs). Such a mapping would enable operators to act on the underlying levers influencing latency-critical KPIs, thereby allowing for targeted and effective latency mitigation. However, the sheer volume of available performance metrics in the Operations Support System (OSS) makes it difficult to manually identify the KPIs most relevant for latency optimization.

With the advent of the Zero-touch Network and Service Management (ZSM) paradigm [14] and advances in Artificial Intelligence (AI), purely analytical models have evolved into data-driven frameworks relying on Supervised Learning (SL) to automatically capture relevant relationships among cell-level measurements and target KPIs. These models can be exploited by MNOs to assess network response under specific network configurations and network conditions for proactive network management. Existing data-driven latency predictive frameworks focus on mean latency [15], [16], [17]. However, in next-generation networks, SLAs for URLLC and IHRLLC services are defined in terms of Latency-Reliability (LR) trade-off, indicating the percentage of packet transmissions that must meet a predefined end-to-end target latency [18]. While mean latency offers limited insight, LR provides a statistical view of performance, enabling a more comprehensive and actionable understanding of latency behavior under varying conditions.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



So far, the estimation of LR has been unexplored despite its critical relevance for MNOs. This may be due to the difficulties of gathering LR KPIs in vendor equipment, since the latency threshold to consider a packet transmission acceptable varies across services [19]. Such data scarcity prevents the training of SL models that directly provide LR estimates.

A second key limitation of existing latency estimation frameworks is their generic design, which overlooks the fact that RRM schemes may treat traffic from distinct classes (e.g., Guaranteed Bit Rate (GBR) vs. best effort) differently. Thus, latency dependencies on network state often vary with service class. In the absence of a deep knowledge of RRM policies, the understanding of these dependencies is extremely valuable for decision-making. While some studies acknowledge the value of reducing the input feature space of SL latency estimation models [15], they lack service-class-specific dimensionality reduction process boosting model efficiency and interpretability.

This work contributes to SL-based latency modeling in mobile networks by addressing key gaps in the literature. It introduces the first hybrid framework that combines analytical modeling and data-driven techniques to estimate LR in the RAN for different service classes, using cell-level KPIs collected from 5G vendor equipment. The target latency is that experienced by a packet from the Data Network (DN) to the User Equipment (UE), as it passes through the Next Generation NodeB (gNB). A full-stack analysis of gNB operations identifies critical delay stages, which are mapped to relevant KPIs collected via the OSS and supports latency modeling through Queueing Theory. Based on the above analyses, a predictive framework is proposed that estimates LR per service class by combining SL, domain knowledge, and statistical latency analysis.

Recall that, in this work, 'prediction' and 'estimation' terms are indistinctly used to refer to the computation of a KPI value at a given time instant from input features reflecting network state at the same time instant.

Thus, the core contributions of this work are threefold:

a) **Analytical latency modeling**: A full-stack analysis of the 5G Quality of Service (QoS) architecture is conducted to trace latency-critical stages in RAN packet transmission. This analysis enables a Queueing Theory-based statistical characterization of latency that, combined with protocol-level insights, reveals dependencies on measurable KPIs. The resulting mapping of delay stages to cell-level KPIs available in OSS data supports both the selection of input features for latency prediction models and the derivation of actionable insights for MNOs.

b) **Data-driven LR estimation**: A novel SL-based framework is proposed to estimate mean packet latency in cellular networks, leveraging the KPIs identified in a). LR metrics are then derived by combining latency estimates with the latency statistical characterization introduced in a). A correlation analysis on real network data motivates the adoption of separate models per service class, with tailored input features and trainable parameters. The best-performing models are selected by comparing different SL approaches and feature sets in terms of accuracy, complexity, and explainability.

c) **Model explainability**: An explanatory analysis with Shapley values is carried out, highlighting the relevance of specific KPIs in estimating mean latency, while revealing unexpected relationships and behaviors due to feature correlation.

The rest of the paper is structured as follows. Section II reviews related work in the literature. Section III presents the network model through a full-stack analysis targeting data flow in a gNB. Section IV formulates the problem of estimating LR in a 5G RAN from cell-level KPIs. Section V presents the proposed LR estimation framework, assessed over a dataset from a live network in Section VI. Finally, Section VII summarizes the main conclusions and outlines future research lines.

## II. Related Work

The following literature review covers the critical role of latency in 5G networks, mathematical models for root cause analysis of latency issues in 5G RAN and data-driven frameworks for radio KPI estimation.

### A. Latency in 5G: importance and mitigation

The critical role of latency for the effective delivery of URLLC services in B5G networks has spurred research on latency mitigation. In [7], Parvez *et al*. categorize low-latency solutions into RAN, Core Network (CN), and caching domains. Likewise, they highlight 5G network elements crucial for meeting stringent latency requirements, such as SDN, NFV, and MEC. Similarly, [20] provides a holistic analysis of design principles and enabling technologies for low-latency networks, emphasizing the trade-offs with other performance metrics and the importance of addressing cross-layer interactions within the communication protocol stack. Correia *et al*. discuss low-latency solutions including MEC node deployment, functionality splitting, radio techniques and network architectures, concluding that MEC is key to achieving end-to-end latency values below 1 ms [21]. In the RAN, most efforts to meet stringent latency requirements have focused on numerology [22]. However, the flexibility it offers often comes at the expense of increased signaling and required spectrum reserved for URLLC [23]. More recently, [24] and [25] showed that UE capabilities can significantly influence latency, even in optimized RAN environments.

In conclusion, while extensive research has proposed various techniques for latency mitigation, many of these approaches tend to overlook practical constraints faced by MNOs, such as the high implementation costs of architectural changes and the lack of scalable business models.

### B. Root-causing latency in 5G networks

Reducing E2E latency in the U-Plane involves optimizing every step of the data delivery process [26]. As a result, several works focus on breaking down the U-Plane Latency components through theoretical analysis [11], [12], [13], [27], [24], [28]. Some recurring elements in these analyses are (i) the time spent waiting for radio resource allocation, (ii) the transmission time over the radio channel, (iii) the delay

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



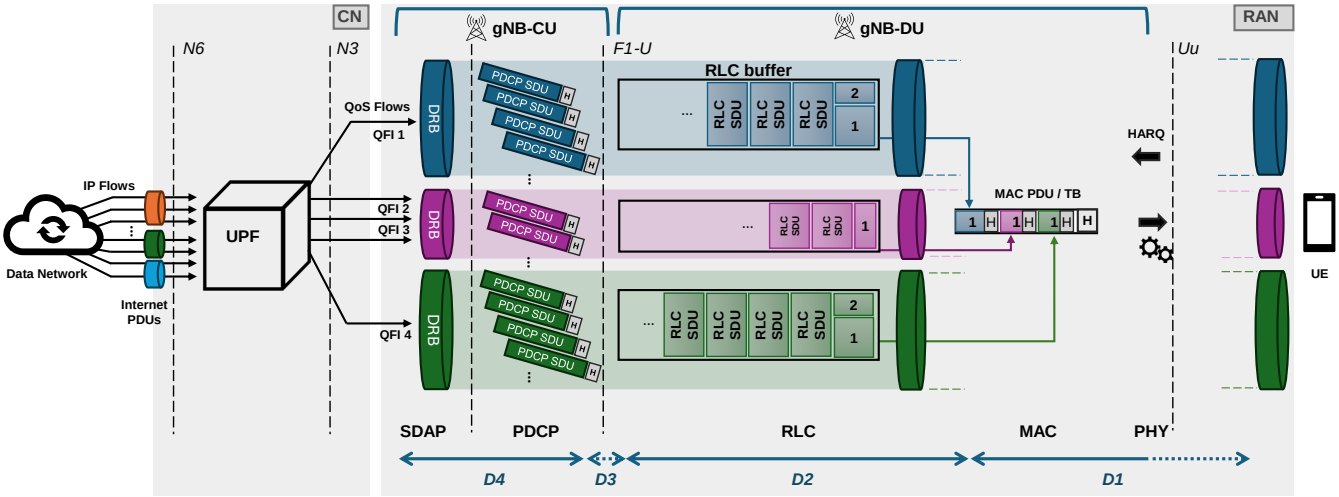



Fig. 1: 5G QoS Architecture - User Plane: DL PDCP SDU transmission at UE level [37].

introduced by pre/post-processing operations to encode/decode information, and (iv) the time required for retransmissions. In some cases, latency formulation accounts for the specific numerology scheme [13] and scheduling and retransmission mechanisms [29].

However, existing analytical latency models are often validated through simulations based on assumptions that cannot reflect the complexity of real network deployments, marked by service mix, heterogeneous RAN topology and proprietary RRM mechanisms. Moreover, these models typically stop at identifying protocol-level delay sources without linking them to observable KPIs, thus offering limited support for targeted actions. In contrast, this work leverages real network data and links latency-critical phases to measurable KPIs, enabling effective, actionable optimization.

### C. ML-based KPI estimation in the RAN

The interest of MNOs for proactive network management has propelled significant research into KPI estimation through SL models trained on network data. SL-based throughput estimation has been extensively covered due to the critical impact of this KPI on data-hungry services. Earliest works relied on Multivariate Linear Regression (MLR) to predict aggregated cell throughput in High-Speed Downlink Packet Access (HSDPA) [30] and multi-service Long Term Evolution (LTE) [31] systems. More advanced SL algorithms were later introduced to capture non-linear relationships between input features and throughput. For instance, [32] employs a deep Artificial Neural Network (ANN) to estimate Downlink (DL) cell/user throughput in a live LTE network, while [33] assesses distance-, vector- and ANN-based algorithms to estimate the same KPIs in HSDPA and LTE, deriving more compact models through Sequential Feature Selection (SFS). With the deployment of live 5G networks, the SL approach has been expanded to address throughput estimation in sliced RANs [34], [35], [36]. Beyond throughput, the estimation of other critical KPIs for 5G services has recently been explored. For instance, [38] uses MLR, decision trees, Support Vector Regression (SVR) and Gaussian process regression to estimate video-streaming key quality indicators from low-layer network measurements, improving slice negotiation. In [39], the challenge of predicting spectrum efficiency in massive multiple-input multiple-output-enabled 5G networks is tackled by leveraging reference signal received power data collected from drive tests that feed a Multi-layer Perceptron (MLP). MLP is also used in [40] to estimate energy consumption of a radio unit from engineering parameters and cell-level KPIs. Similarly, data-driven techniques have been employed for broader RAN optimization tasks. For instance, in [41], a Deep Reinforcement Learning (DRL)-based Remote Radio Head (RRH)–Baseband Unit (BBU) mapping scheme is proposed to enhance performance in large-scale Cloud-RANs.

Latency estimation is especially challenging due to its non-linear relationship with highly variable factors, such as network congestion, dynamic routing, buffer management (e.g., queuing delays, packet drops...) and packet scheduling. In fact, [42] proves that, while MLR can estimate cell throughput in a multi-service LTE network reasonably well, it struggles to predict packet delay for Voice over Internet Protocol (VoIP) users. Among studies employing data-driven SL models for latency prediction, [43] lays the groundwork by estimating Round Trip Time (RTT) for PING messages to LTE servers using Logistic Regression, SVR and Decision Trees, leveraging real-world network data. [23] and [17] focus on predicting extreme latency events, which are particularly critical in environments with stringent reliability requirements. In [23], synthetic data is used to predict latency degradation in industrial plants through classification and regression tasks with ANNs and k-Nearest Neighbors (KNN), highlighting the challenge of imbalanced datasets due to rare events. [17] estimates the tail of packet latency distribution with a hybrid model combining Gaussian mixture models for bulk latency and generalized Pareto distributions for extreme values. This approach leverages real data from Wi-Fi, commercial private and software-defined 5G networks to predict the likelihood of rare latency events based on network conditions. Unlike our approach, [17] adopts a generic framework, overlooking service-class-specific KPIs and their varying relevance, without tailoring

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



predictions through targeted KPI selection. Closer to this work, [15] predicts packet latency for Voice over LTE (VoLTE) service using MLP, SVR and KNN, highlighting the critical role of minimizing the number of predictors in SL models to enhance performance. However, ground-truth latency values are not directly measured, but derived from a polynomial function relating latency to the number of users, which may limit model generalizability due to the strong influence of packet scheduling mechanisms in such a relationship. Finally, in [16], mean cell-level latency for services with QoS Class Identifier (QCI) 1 and 7 is estimated with a Probabilistic Bayesian Neural Network (BPNN) trained on data gathered in a live 4G network. Nonetheless, none of these contributions provides a framework for estimating LR per service class in B5G networks.

Overall, while prior work has demonstrated the potential of SL-based models for KPI estimation, including latency, existing approaches often rely on synthetic or indirect ground-truth measurements, and overlook the differentiated impact of service classes, thus not explicitly addressing LR estimation on a per-class basis within B5G networks.

## III. Network Model

In this section, the network model is presented by conducting a full-stack analysis within the gNB, focusing on the flow of data packets arriving from the DN to the gNB and traversing the entire protocol stack until they are successfully delivered to the UE, as illustrated in Fig. 1. Throughout the description, particular attention is given to identifying the processes that contribute most significantly to latency.

### A. 5G QoS architecture

The inherent heterogeneity of 5G use cases creates a complex QoS scenario. To ensure that data packets meet specific service requirements, QoS flows are used [37]. Upon entering the 5G system, Internet Protocol (IP) packets are grouped into QoS flows by the Non-Access Stratum (NAS) in the CN. Each flow is assigned a QoS Flow Identifier (QFI) by the Packet Detection Rule (PDR) algorithm at the User Plane Function (UPF). The QFI is a 6-bit field determining the packet forwarding treatment across the entire Packet Data Unit (PDU) session, i.e., the logical connection established between the DN and the UE [37]. As a result, each QoS flow is linked with a QoS profile defined by several parameters, such as Allocation Retention Priority (ARP) and 5G QoS Identifier (5QI). The latter, analogous to QCI used in 4th Generation (4G), is a scalar that points to a set of QoS characteristics, including maximum data burst volume, resource type (GBR, non-GBR, delay-critical GBR), packet priority level, and tolerable Packet Delay Budget (PDB) and packet error rate.

Then, in the RAN, Service Data Adaptation Protocol (SDAP) layer maps QoS flows to Data Radio Bearers (DRBs) (different colors in Fig. 1) at the gNB-Centralized Unit (CU) level. Since only 30 DRBs are available per UE, a many-to-one relation occurs, generating a funnel with limited ability to segregate packets when the number of flows grows [44].

### B. IP packets flow: path to RLC buffers

Following the SDAP layer in the U-Plane of the gNB-CU, the Packet Data Convergence Protocol (PDCP) layer is responsible for header compression, security measures such as integrity protection and ciphering, as well as re-ordering and retransmission during handovers. Once PDCP headers are appended to IP packets, PDCP PDUs are passed to the underlying Radio Link Control (RLC) sublayer, reaching the gNB-Distributed Unit (DU) via the F1-U interface. Packets carried through the same DRB are queued in separate RLC buffers. As packets in a buffer belong to the same QoS flow, they are retrieved according to a First-In, First-Out (FIFO) policy to ensure equal prioritization. An adverse effect on packet delays at the RLC layer is bufferbloat. This phenomenon occurs when RLC buffers are configured with large capacities to manage fluctuations in radio channel availability. The network stores data during periods of high channel capacity and releases it when capacity decreases, aiming to maintain consistent performance. However, while this approach ensures smooth data flow, it can mislead Transmission Control Protocol (TCP) congestion control by masking packet losses, causing TCP to increase the transmission rate, exacerbating congestion and delays erroneously [44].

RLC sublayer supports Transparent Mode (TM), Unacknowledged Mode (UM), and Acknowledged Mode (AM) modes [45]. AM employs Automatic Repeat request (ARQ) with Acknowledges (ACKs) and retransmissions, and is typically used for packets from services where reliability is crucial but latency is not critical. Since this work focuses on low/critical-latency services, UM mode is assumed hereafter, so that error correction is exclusively handled at the Medium Access Control (MAC) layer.

### C. Interplay between RLC and MAC sublayers

The underlying MAC sublayer is responsible for packet scheduling. Recall that, at the gNB, each UE has multiple RLC buffers (one per DRB) containing queued RLC Service Data Units (SDUs) awaiting DL transmission. Packets within the same buffer share the same 5QI value. Every Transmission Time Interval (TTI), the scheduler determines how many RLC SDUs to retrieve from each buffer to fill a Transport Block (TB), transmitted to the UE through the Uu air interface [46]. Once selected, RLC SDUs are multiplexed by the MAC sublayer into a MAC PDU, where each SDU is preceded by a sub-header indicating the logical channel identifier (LCID), its length, and control information. Optional MAC Control Elements (CE) may also be included. The assembled MAC PDU is then encapsulated into a TB for physical transmission [47]. Final delivery to the UE depends on dynamic scheduling decisions, which govern when and how the TB is transmitted over the air interface.

The packet scheduling strategy is vendor-specific, hence unknown to the MNO. Nonetheless, the main stages in a generalized QoS/channel-aware packet scheduler can be outlined to identify delays causes within MAC sublayer. The packet scheduler unit comprises a Time Domain (TD) scheduler and a Frequency Domain (FD) scheduler. The TD scheduler

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



prioritizes data flows based on the service class (i.e., 5QI) and other parameters (e.g., PDB, ARP). It also monitors the RLC buffer status and considers factors related to traffic volume to ensure decisions align with current network conditions and demand. Subsequently, the FD scheduler allocates specific Physical Resource Blocks (PRBs) for DL TB transmission toward the UE. The selection of PRBs is driven by the UE instantaneous channel quality, periodically reported through Channel Quality Information (CQI) reports. These reports provide channel quality information for each sub-band within the Channel State Information (CSI) reporting band. Each CQI value is associated with a specific modulation scheme, ranging from Binary Phase Shift Keying (BPSK) to 256-Quadrature Amplitude Modulation (QAM), along with a corresponding coding rate, which together define channel capacity and spectral efficiency. Although the CQI value suggests an appropriate Modulation and Coding Scheme (MCS) to maintain a predefined Bit Error Rate (BER), the final MCS selection is determined by the Link Adaptation Unit (LAU) using an Adaptive Modulation and Coding Scheme (AMCS) approach, which considers the specific MCS supported by the UE along with the QoS requirements. Thus, QoS awareness is managed by the TD scheduler, while radio-channel awareness is handled by the FD scheduler.

Delay in MAC scheduling is impacted by three processes:

**1) UE awareness of DL Resource Allocation.** Before decoding data on the Physical Downlink Shared Channel (PDSCH), each UE scans the Physical Downlink Control Channel (PDCCH) to read the Downlink Control Information (DCI) associated with its Cell Radio Network Temporary Identifier (C-RNTI), obtaining details on resource allocation, modulation and coding scheme, Transport Block Size (TBS), and Hybrid Automatic Repeat request (HARQ) parameters. DCI resides within the first three Orthogonal Frequency-Division Multiplexing (OFDM) symbols of the subframe in one TTI, with the remaining symbols dedicated to data transmission. Thus, an RLC SDU remains in the RLC buffer until it is scheduled. Once scheduled, the UE aligns with the subframe to scan the PDCCH and read the DCI, and finally accesses the scheduled PRBs to receive DL data. Insufficient PDCCH resources can delay DL scheduling.

**2) RLC segmentation.** At the transmitter side, if necessary, the MAC sublayer instructs the RLC sublayer to segment the RLC SDU to fit within the TB. As a result, only part of the PDCP SDU is transmitted within a single TTI. While this optimizes spectrum usage when packet sizes exceed available link capacity, it can also introduce significant delays, as all packet segments must be reassembled at the receiver before being forwarded to the upper layers. Moreover, any delay or loss of a segment can hold back the entire packet, increasing latency.

**3) HARQ operation.** In 5G New Radio (NR), 16 several parallel HARQ processes ensure error-free data transmission in the MAC layer. As a consequence, multiple TBs can be sent concurrently, even before receiving ACK/Negative Acknowledgment (NACK) feedback for previously transmitted TBs. Each HARQ process operates in stop-and-wait mode, with an associated HARQ ID that allows the UE to inform the gNB whether a TB is successfully received or requires retransmission. This approach allows for out-of-sequence data delivery, helping to minimize latency. The HARQ RTT for a single TB varies based on the selected numerology. If a retransmission is required, the RTT doubles with each subsequent attempt. The scheduling of retransmissions is handled by the HARQ manager at the MAC layer. Consequently, the sojourn time in the RLC buffer also depends on whether the scheduler prioritizes retransmissions or new transmissions.

## IV. Problem Formulation

This section outlines the LR estimation problem in a 5G RAN based on data collected by the OSS. To this end, the problem structure is first presented and the LR metric is formally defined. Then, latency metric proposed by 3rd Generation Partnership Project (3GPP) is compared with that implemented by vendors. Finally, the probability distribution associated with vendor-specific latency measurements is derived through Queueing Theory.

### A. Latency-reliability estimation problem

In this work, the estimation of LR is formulated as a regression problem tackled through SL. The aim is to develop a SL model that predicts LR as the target variable, from a set of cell-level KPIs, identified through the full-stack analysis detailed in Section III. The optimization goal is to minimize the prediction error with respect to a defined Figure of Merit (FoM). Machine Learning (ML), and specifically SL, is adopted instead of traditional analytical approaches due to its ability to capture complex interactions among KPIs that are hard to model explicitly. Moreover, unlike analytical models, which rely on scenario-specific assumptions and manual tuning, SL-based frameworks are flexible and scalable, allowing generalization across cells and services with different RRM configurations and deployment conditions without the need for model redesign. To understand the adopted resolution strategy, the concept of LR is first formalized.

The 3GPP TS 22.261 specification defines reliability as the percentage of packets successfully delivered to a system entity within the required time constraint, out of all packets transmitted. Mathematically, this is represented as

$$LR = 100 \cdot \frac{N_p(L_p < L^{(max)})}{N_p}, \tag{1}$$

where $N_p$ is the total number of packets transmitted in the cell, and $N_p(L_p < L^{(max)})$ is the number of packets received with a packet latency, $L_p$, below a predefined threshold, $L^{(max)}$, set according to the latency demand associated with the target service class (QCI). From (1), it can be noted that estimating LR requires deriving the statistical distribution (Probability Density Function (PDF)) of the packet latency measured through vendor-specific RAN equipment. To this end, the vendor-specific latency metric is compared with the one defined by 3GPP, enabling its proper characterization and subsequent modeling for PDF derivation.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



### B. Latency definitions

The 3GPP Technical Report (TR) 38.913 defines U-Plane Latency as "the one-way transit time from when a packet is available at the IP layer of the UE (Uplink (UL)), or gNB (DL), to when it reaches the corresponding IP layer on the receiving side, including any necessary HARQ retransmissions" [10]. This definition is primarily theoretical and serves as a reference for setting the latency requirement under ideal conditions, i.e., small IP packets (e.g., 0-byte payload), no discontinuous reception (DRX), negligible scheduling delays [48], etc. This latency, hereafter referred to as $L_{\text{3GPP}}$, is the sum of many delays introduced by the different processes involved in data transmission through the air interface explained in Section III. In Technical Specification (TS) 38.314, four packet delay measures for the DL RAN are introduced, namely *D1*, *D2*, *D3* and *D4* (shown in Fig. 1). These measures, defined in TS 28.552, are described next:

- ***D4*** represents the time duration within the gNB-CU U-Plane, encompassing the interval from when an IP packet is received to when the corresponding PDCP SDU is transmitted over the F1-U interface.
- ***D3*** quantifies the average DL packet delay across the F1-U interface, capturing the transmission time between the gNB-CU U-Plane and the gNB-DU.
- ***D2*** is the delay experienced at the RLC layer, defined as the time elapsed from when an RLC SDU is enqueued in the RLC buffer until the last segment of the RLC SDU is scheduled for transmission by the MAC layer and subsequently removed from the buffer for inclusion in a TB. Since the comprehensive full-stack analysis has identified the main bottlenecks within the RAN between the PDCP and MAC layers, *D2* is considered the most critical component in RAN latency.
- ***D1*** captures the delay from the delivery of the last segment of an RLC SDU to the UE until it is acknowledged by the MAC layer via HARQ feedback, including potential retransmissions.

Currently, 5G network equipment provided by vendors computes latency on a per-packet basis, measuring the time from when a packet arrives at the gNB to the transmission of its first segment over the air interface [49]. This latency metric, hereafter referred to as $L_{\text{vendor}}$, includes delay components *D4*, *D3* and a subset of *D2*, denoted as *D2'*, which captures the delay until the first segment of the RLC SDU is inserted into a TB. This approach is motivated by the substantial processing time and computational inefficiency associated with measuring $L_{\text{3GPP}}$ (i.e., the full latency from the receipt of a PDCP SDU to its complete transmission with HARQ loop) using current tools [49]. $L_{\text{vendor}}$ is often averaged across all packets transmitted in the DL during a Reporting Period (ROP) (e.g., 15, 30, or 60 minutes [50]), with aggregation performed separately per 5QI to account for service-specific latency requirements. Resulting KPIs characterize cell performance per service type independently of packet length, and thus it is an effective metric for specifying LR requirements in SLA for network slices delivering services with variable burst sizes (e.g., URLLC slices supporting autonomous driving or extended reality applications). In applications with small packet sizes, such as tactile Internet, autonomous driving, or conversational voice, entire PDCP SDU are often transmitted within a single TTI, leading *D2* to coincide with *D2'*. Moreover, when retransmissions are omitted to minimize delay, *D1* reduces to a constant delay from the HARQ feedback, not impacting the shape of latency PDF. Thus, $L_{\text{vendor}}$ allows for computational feasibility while capturing key latency characteristics in practical deployments.

### C. Mathematical characterization of latency

Once clarified the nature of the observed latency metric, the derivation of its statistical distribution is addressed. In this work, $L_{\text{vendor}}$ is modeled using Queuing Theory. For generality, the latency to transmit an entire PDCP SDU in the DL of a 5G network (i.e., $L_{\text{3GPP}}$ excluding the HARQ loop and potential retransmissions) is first derived. Then, $L_{\text{vendor}}$, representing the latency up to the first segment transmission, is treated as a specific case of the previous one. Note that *D3*=0 in the absence of a functional split, while *D4* is negligible compared to *D2* (microseconds vs. milliseconds). Thus, the analysis focuses on the time the PDCP SDU spends in the RLC buffer and its subsequent transmission.

For a simpler analysis, for a given UE $u$, each RLC buffer is modeled as an $M/M/c$ queue, where PDCP SDUs of a given QoS flow arrive following a Poisson process with rate $\lambda$ and are processed by $c$ independent servers whose service time is exponentially distributed with rate $\mu$ [51]. This queueing model assumes an infinite RLC buffer size, thus neglecting the possibility of packet losses due to buffer overflow. $\lambda$, representing the IP packet arrival rate, depends on the service type accommodated in the corresponding DRB (e.g., full-buffer services lead to higher $\lambda$ than conversational voice service). Likewise, $\mu$, representing the inverse of the average time required to transmit a complete PDCP SDU once the first segment has been scheduled, depends on the extent of segmentation, determining the number of TTIs needed per SDU. Even though the physical-layer channel is not explicitly modeled, $\mu$ implicitly captures its effects, as it varies with cell configuration (e.g., cell bandwidth, numerology, etc.), channel conditions, cell load, scheduling policy and packet size. Consequently, $\mu$ can be modeled as a stochastic variable. Finally, $c$, representing the number of parallel servers, can be interpreted as the average number of RLC SDUs (or segments thereof) that the scheduler can simultaneously retrieve from each RLC FIFO buffer within a single TTI. The full-stack analysis in Section III highlighted several factors influencing the value of $c$ under a QoS/channel-aware packet scheduling policy. These factors include queue priority indicated by 5QI, current traffic conditions, the volume of retransmissions, PDSCH load given by PRBs utilization and PDCCH load. Consequently, the value of $c$ may vary for each RLC buffer. The sojourn time of a packet in the system can be computed as the sum of the time spent in the buffer (queue), $T_q$, plus the service time, $T_s$. Thus, the PDF of $L_{\text{3GPP}}$ can be computed as $f_{L_{\text{3GPP}}}(t) = f_{T_q+T_s}(t)$. Waiting time probability distribution can be derived by differentiating its Cumulative Distribution

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



Function (CDF), $F_q(t)$. Specifically, for an $M/M/c$ system with a FIFO queueing policy, $f_q(t)$ is given by [52]

$$\begin{aligned} f_q(t) &= \frac{d}{dt}F_q(t) = \frac{d}{dt}\left(1 - e^{-(c\mu-\lambda)t}\right) \\ &= \underbrace{(c\mu - \lambda)}_{k} e^{-\overbrace{(c\mu-\lambda)}^{k}t}. \end{aligned} \tag{2}$$

By denoting $k = c\mu - \lambda$, it can be noticed that waiting time is exponentially distributed. Likewise, service time distribution is given by $f_s(t) = \mu e^{-\mu t}$. By assuming that $T_q$ and $T_s$ are independent, the total latency distribution can be computed through the convolution of the two PDFs as

$$\begin{aligned} f_{L_{\text{3GPP}}}(t) &= (f_q * f_s)(t) = \int_0^t f_q(\tau) \cdot f_s(t-\tau)\, d\tau \\ &= \int_0^t \left(k \cdot e^{-k\tau}\right)\left(\mu e^{-\mu(t-\tau)}\right) d\tau \\ &= \frac{k\mu}{k-\mu}\left(e^{-\mu t} - e^{-kt}\right). \end{aligned} \tag{3}$$

For the specific case of $L_{\text{vendor}}(t)$, the focus is on calculating the sojourn time of only the first segment of a PDCP SDU. Regarding $f_s(t)$, since $T_s$ corresponds to the interval between the arrival of the PDCP SDU in the buffer and the scheduling of its first segment, the waiting time experienced by the entire PDCP SDU coincides with that of its first segment. As for the service time, while the entire PDCP SDU is subject to a variable service time (with $T_s$ exponentially distributed), the time required to serve only the first segment (or any other) once scheduled is deterministic and corresponds to the TTI duration, given by the selected 5G numerology as specified in [53] (e.g., in this work, $\nu = 0$ is considered, corresponding to a TTI duration of 1 ms). Consequently, the system is simplified to an $M/D/c$ model, with $c$ the average number of PDCP SDUs in a RLC buffer whose first segment is transmitted in the same TTI. In this system, the latency PDF follows the exponential distribution specified in (2), shifted by $T_s$, i.e.,

$$f_{L_{\text{vendor}}}(t) = \begin{cases} (c\mu - \lambda) \cdot e^{-(c\mu-\lambda)(t-T_s)}, & t \geq T_s \\ 0, & t < T_s. \end{cases} \tag{4}$$

Deriving the PDF in (4) requires the estimation of $\lambda$ and $c$ in the $M/D/c$ model. However, this task is extremely challenging due to the dependence of $\lambda$ and $c$ on network conditions (e.g., traffic load and channel quality), service class characteristics (e.g., QCI/5QI), and cell configuration parameters (e.g., 5G numerology, cell bandwidth, packet scheduler, etc.). Nonetheless, note that $f_{L_{\text{vendor}}}(t)$ follows an exponential distribution characterized by rate $k = c\mu - \lambda = 1/\overline{f_{L_{\text{vendor}}}}(t)$, with $\overline{f_{L_{\text{vendor}}}}(t)$ the mean packet latency. Since the mean fully characterizes the exponential distribution, LR can be inferred by evaluating $f_{L_{\text{vendor}}}(t)$ over an estimate of the mean latency for each service class, cell, and ROP. This approach is adopted in this work, where the mean latency estimation ($\widehat{\overline{L_{\text{vendor}}}}$) is performed through Bayesian supervised learning using cell-level KPIs selected from the full-stack analysis presented in Section III.

## V. ML-Empowered Latency Predictive Framework

This section introduces a novel framework for estimating LR in cellular RANs. Fig. 2 outlines the workflow of the proposed system, integrating SL with an analytical latency model. The process includes three stages. First, vendor KPIs affecting latency according to the above-presented full-stack system analysis and mathematical latency characterization are identified. These are gathered by the live network on a cell and ROP basis, and later pre-processed (e.g., data normalization, dataset split) to train a SL model. Then, a Bayesian Neural Network (BNN) is used to estimate mean latency per cell, service class (i.e., QCI/5QI) and ROP. To enhance efficiency, dimensionality reduction is performed through SFS. For model explainability, a Shapley analysis is performed. Finally, LR is estimated from mean latency by using the latency PDF obtained in Section IV. LR estimates provided by the proposed system can be used for decision-making in various network optimization tools (e.g., traffic steering algorithms or cell issue detectors) to proactively find the RAN configuration that ensures SLA fulfillment. All these steps are detailed next. Then, the theoretical time complexity of the proposed framework is discussed.

### A. Data preparation

The input data for the predictive framework comprises Configuration Managements (CMs) and Performance Measurements (PMs) collected by current vendor equipment at cell level on a ROP basis. These raw measurements are gathered in the OSS and processed to derive cell-level KPIs, offering actionable insights for network monitoring and optimization.

Table I presents the set of features (KPIs) considered in this work, detailing the name, description, type (C=Channel, T=Traffic, CAP=capacity, O=output) and related protocol layer (if any), the latter aimed at assisting network operators in identifying and optimizing specific aspects of network performance. Blank rows contain candidate predictors for the SL model used to estimate mean latency. These 22 input features are related to processes affecting RAN latency identified through the full-stack analysis and latency mathematical characterization discussed above. For instance, *block_pdcch* relates to the process where insufficient PDCCH resources delay DL scheduling. *rtx_vol*, *pdcp_qci1*, *pdcp_qci7*, and *prb_util* impact the packet sojourn time, determined by current traffic conditions and composed of buffer time $T_q$ and service time $T_s$. The latter depends on the TB size (*tb_size*), also included as a predictor. Light-grey rows include the mean packet latency in DL for QCI1 (VoLTE) and QCI7 (live video streaming, interactive gaming, etc.), used as ground truth. Hereafter, these KPIs are denoted as $\overline{L_{\text{QCI1}}}$ and $\overline{L_{\text{QCI7}}}$, whereas $\overline{L_{\text{QCI}X}}$ is used to refer to a generic QCI. $\overline{L_{\text{QCI}X}}$ is equivalent to $L_{\text{vendor}}$, measuring the mean retention period of PDCP SDUs at the gNB. Finally, dark-grey row corresponds to a KPI denoting the percentage of QCI1 packets whose retention period plus HARQ transmission time is below 100 ms. Such a KPI is used to assess the effectiveness of the proposed framework in estimating LR. Note that, in the framework exploitation stage, only input features must be collected.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



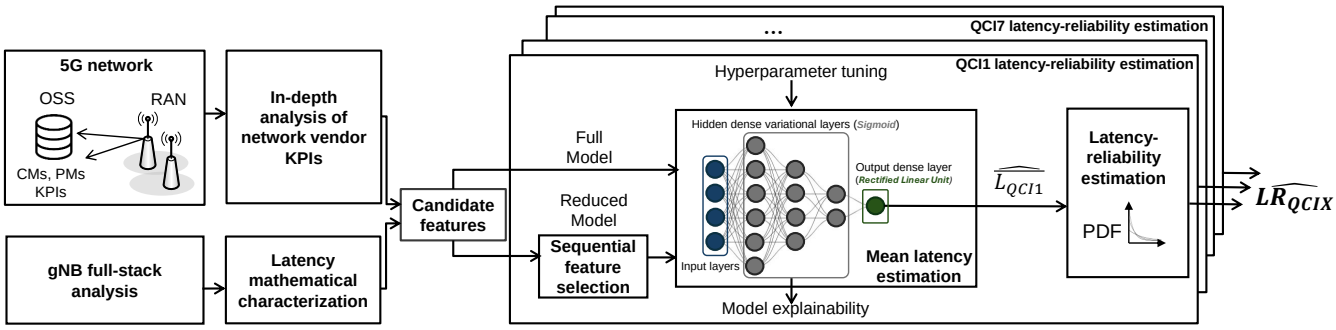



Fig. 2: Workflow of the proposed ML-empowered latency-reliability predictive framework.

TABLE I: Vendor KPIs used to estimate mean latency.

| KPI name | Description | Type | Layer | SFS QCI1 | SFS QCI7 |
|---|---|---|---|---|---|
| *ue_qci1* | No. active UEs QCI1 in DL | T | | ✓ | |
| *ue_qci7* | No. active UEs QCI7 in DL | T | | | ✓ |
| *ue_tot* | No. total active UEs in DL | T | | ✓ | ✓ |
| *ue_sched* | % scheduled vs active UEs in DL | T | MAC | | |
| *avg_buff* | Avg. buffer size for DL | CAP | RLC | | ✓ |
| *max_buff* | Max. buffer size for DL | CAP | RLC | | ✓ |
| *block_pdcch* | No. blocked attempts on PDCCH | T | MAC | | ✓ |
| *prb_util* | PRB utilization ratio in DL | CAP | | ✓ | ✓ |
| *available_prb* | Mean PRBs available in DL | CAP | | | ✓ |
| *pdcp_qci1* | No. PDCP SDUs tx in DL for QCI1 | T | MAC | ✓ | |
| *pdcp_qci7* | No. PDCP SDUs tx in DL for QCI7 | T | MAC | | ✓ |
| *rlc_dl* | No. RLC SDUs tx in DL | T | RLC | ✓ | ✓ |
| *rtx_vol* | % retx traffic volume | T | MAC | | |
| *tot_vol* | Total traffic volume [bytes] | T | RLC | | |
| *mcs_dl* | Avg. MCS in DL | C | MAC | | ✓ |
| *mcs_ul* | Avg. MCS in UL | C | MAC | | |
| *avg_cqi* | Avg. CQI | C | MAC | | ✓ |
| *dl_se* | DL Spectral Efficiency [$\frac{bits}{Hz \cdot s}$] | C | MAC | | |
| *low_mcs_tx* | % tx using low MCS in DL | C | MAC | | ✓ |
| *high_mcs_tx* | % tx using high MCS in DL | C | MAC | | |
| *tb_size* | Avg. TB size in DL [bytes] | CAP | MAC | | ✓ |
| *rrc_conn* | No. RRC connections | T | | ✓ | ✓ |
| $\overline{L_{\text{QCI1}}}$ | Mean Latency QCI1 in DL [ms] | O | | | |
| $\overline{L_{\text{QCI7}}}$ | Mean Latency QCI7 in DL [ms] | O | | | |
| $LR_{\text{QCI1}}$ | DL Latency-reliability QCI1 [%] | O | | | |

To ensure high accuracy and faster convergence of SL models, input features are standardized using $Z$-score normalization, which guarantees that each feature distribution across datapoints has 0 mean and unit variance. To avoid cell and/or time correlation among consecutive training samples, the dataset is randomly shuffled. Then, data is split into training and testing sets.

### *B. Mean latency estimation*

The aim of the ML-based framework is to predict LR from network data, which requires predicting latency distribution per cell and ROP. Among existing SL models, BPNNs are generally used for this purpose, since they model uncertainty and output the target variable distribution rather than a single-point estimate. BPNNs are well-suited for predicting distributions governed by multiple unknown parameters, as the final layer outputs a probabilistic distribution parameterized based on the output of the preceding layers. To train BPNNs, Negative log-likelihood (NLL) is often used as a loss metric to generate output distributions that best fit the observed data [54]. However, the specific characteristics of latency distribution and the available data for model training make BPNNs unsuitable for the purpose targeted here. As outlined in Section IV-C, RAN packet latency follows an exponential distribution characterized uniquely by a rate. When ground truth exclusively comprises mean value observations, NLL tends to be dominated by datapoints with high rate (i.e., low mean latency), leading to underestimation of higher latency instances. This results in predictions that focus on the central trend while failing to account for worst-case scenarios where latency increases considerably. To overcome this issue, recall that the rate parameter of an exponential distribution is the reciprocal of the mean. Consequently, since $\overline{L_{\text{QCI}X}}$ measurements are available in current cellular networks, the task of predicting latency distribution reduces to a regression problem focused on estimating the mean, which fully defines the PDF. In the proposed predictive framework, $\overline{L_{\text{QCI1}}}$ and $\overline{L_{\text{QCI7}}}$ are estimated by using a BNN (a separate BNN per QCI) fed with the set of input features introduced in Table I. Unlike traditional regression approaches, such as distance-based methods (e.g., K-Nearest Neighbors (K-NN)), vector-based models (e.g., linear regression, SVR), and tree-based algorithms (e.g., decision trees, random forests, boosting), BNNs can model complex non-linear patterns, while capturing model and data uncertainty. The regression model architecture and dimensionality reduction strategy are detailed next.

**1) Regression model.** Fig. 2 outlines the architecture of the BNN. The input layer receives the feature vector and passes it through several variational dense layers that model weights and biases as probability distributions. These hidden layers then apply a *Sigmoid* activation function to introduce non-linearity, allowing the network to capture complex patterns in data. The output layer is a dense layer with one neuron followed by a rectified linear activation function (since latency cannot take negative values). In the variational layers, the model assumes an isotropic Gaussian prior distribution over trainable parameters, serving as a regularizer to control model complexity. The posterior distribution, also Gaussian, is learned during training, where the mean ($\alpha$) and variance ($\sigma^2$) are the trainable parameters that capture the uncertainty in weights and biases. This probabilistic formulation captures epistemic uncertainty, i.e., the uncertainty over model parameters arising from limited or noisy training data. In 5G RANs, where latency is influenced by dynamic and partially observable conditions, this allows the model to express lower confidence on rare or unseen traffic patterns, providing oper-

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



ators with insight into the reliability of predictions.

Training is driven by backpropagation to minimize the Root Mean Squared Error ($RMSE$), defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{5}$$

where $N$ is the number of training datapoints, and $y_i$ and $\hat{y}_i$ are the actual and predicted value of the target variable, respectively. Additionally, the Kullback-Leibler (KL) divergence, given by $\mathrm{KL}(q(\theta)||p(\theta)) = \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta$ [54], is incorporated by the variational layers to regularize the model by constraining the learned posterior $q(\theta)$ toward the prior $p(\theta)$, mitigating overfitting. This allows the BNN to balance prediction accuracy with uncertainty modeling. At inference time, the BNN uses Monte-Carlo sampling by performing multiple forward passes with weights drawn from the learned posterior distribution. Averaging the resulting latency estimates yields the final mean latency prediction.

**2) Dimensionality reduction.** SFS is applied to identify the most relevant predictors to estimate mean latency for each QCI from those initially listed in Table I. SFS starts with an empty model. Then, at each training step, it iteratively adds to the model the input feature contributing most significantly to minimizing a certain loss metric compared to the previous model. This process is repeated until the model considers the full set of candidate input features. SFS enhances computational efficiency by simplifying the model. Moreover, it retains only the most informative features, simplifying model explainability, and reducing data storage and processing capacity required in the OSS. In the proposed framework, SFS process is driven by the Mean Absolute Error ($MAE$), computed as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \ . \tag{6}$$

For optimal model performance, hyperparameter optimization is carried out. To reduce training time, the learning rate, number of hidden layers, neurons per layer, and batch size are tuned at each step of the SFS process through a random grid search with $k$-fold cross-validation. The best hyperparameter tuple is selected as the one minimizing the $MAE$. The impact of each hyperparameter on model performance was also evaluated by analyzing the variation in $MAE$ across their respective ranges. Results show that the learning rate and batch size exhibit limited sensitivity (normalized impact on $MAE < 0.15$), while the number of hidden layers and neurons per layer have a greater influence (normalized impact $\approx$ 0.45 and 0.55, respectively), highlighting the critical role of architectural parameters in capturing non-linear patterns. At the end of the process, each SFS model (i.e., each feature subset with its optimized hyperparameter tuple) is also evaluated in terms of the 95$^{\text{th}}$ percentile of the absolute error, $p_{AE}^{95}$. The best model is then selected by jointly considering both $MAE$ and $p_{AE}^{95}$, enabling a trade-off between average accuracy and robustness under worst-case cell-ROP latency conditions.

MNOs often measure the cost of performance estimation models in terms of model size, computational efficiency, and data handling overhead. SFS is particularly effective in this regard, as reduced models retain comparable prediction performance while significantly lowering the number of input predictors. This simplification leads to faster training and inference times, reduced memory usage, and lower energy consumption, critical in real-time applications. Fewer and interpretable features also enhance model explainability, helping MNOs identify the most relevant KPIs and act on underlying network levers for latency mitigation. Moreover, SFS reduces KPI collection, storage, and processing overhead, supporting more sustainable network operations in large-scale scenarios.

### C. Latency-reliability computation

The exponential latency PDF is characterized by the rate parameter $r = 1/(\widehat{L_{\mathrm{QCI}X}})$. Recall that $\widehat{L_{\mathrm{QCI}X}}$ refers to the time a PDCP SDU spends at the gNB, from reception to the insertion of its first segment into the TB. Therefore, since the total RAN latency must also account for the first segment's transmission time over the radio channel, the exponential distribution is shifted by a constant $T_s$ representing the service time, as in (4), thus obtaining $\widehat{f_{L_{\mathrm{QCI}X}}(t)}$. In LTE, $T_s = 1$ ms. In NR, $T_s$ depends on the numerology (e.g., 1 ms for $\mu = 0$, 0.5 ms for $\mu = 0.5$, etc. [45]). Once the PDF is established, $LR$ is calculated by integrating $\widehat{f_{L_{\mathrm{QCI}X}}(t)}$ up to a predefined latency threshold, $L_{\mathrm{QCI}X}^{(max)}$, which corresponds to the latency requirements of the specific service class. The final expression of the estimated $LR$ for service class QCI$X$ is then given by

$$\widehat{LR_{\mathrm{QCI}X}} = \int_0^{L_{\mathrm{QCI}X}^{(max)}} \widehat{f_{L_{\mathrm{QCI}X}}(t)}\, dt \ . \tag{7}$$

This integral yields the probability that latency experienced by the PDCP SDU remains below the threshold $L_{\mathrm{QCI}X}^{(max)}$, thereby assessing whether the LR requirement for the corresponding service class is met.

Algorithm 1 provides a comprehensive overview of the proposed hybrid mathematical and data-driven framework, emphasizing how the derived equations are employed to compute LR.

### D. Time complexity

Implementing the proposed predictive latency framework in a cellular RAN entails gathering and pre-processing data in the OSS, and training the SL models to estimate mean packet latency per cell and QCI/5QI. Data collection should not pose any additional effort for MNOs, since input features considered in this work correspond to KPIs currently provided by vendor equipment on a ROP basis for network management purposes. Thus, the most time-consuming task is setting and training the optimal SL model, i.e., combination of BNN hyperpameters and set of input features. Such a process must be carried out separately for each specific network and QCI to effectively capture the impact of RRM algorithms on latency. Specifically, the analytical worst-case time complexity of a BNN is $\mathcal{O}\big(\sum_{l=1}^{N_l} N_i(l) \times N_n(l)\big)$, with $N_l$ the number of layers, $N_i(l)$ the number of layer inputs and $N_n(l)$ the number of neurons.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



**Algorithm 1:** Hybrid framework for LR estimation.

**Input:** Cell-level KPIs including latency ($\overline{L_{\text{QCI}X}}$)
**Output:** $\widehat{LR}_{\text{QCI}X}$ estimates, $\theta_{\text{QCI}X}$ for all QCIs
**Prerequisites:**
Eq. (1): LR defined from latency PDF
Eqs. (2)-(4): latency modeled with exponential PDF
**Procedure:**
Extract relevant KPIs for latency from full-stack analysis and queueing model parameters in Eq. (4)
**for** *each QCI class* $X$ **do**
  Initialize: $\mathcal{F}_{\text{sel}} \leftarrow \emptyset$, $\mathcal{F}_{\text{cand}} \leftarrow \{22 \text{ KPIs}\}$
  **while** $\mathcal{F}_{cand} \neq \emptyset$ **do**
    **for** *each* $f \in \mathcal{F}_{cand}$ **do**
      Train a BNN model using features $\mathcal{F}_{\text{sel}} \cup \{f\}$
      Evaluate the model using $MAE$ (Eq. (6))
    **end**
    Select best feature: $f^* \leftarrow \arg\min_f MAE$
    Update sets: $\mathcal{F}_{\text{sel}} \leftarrow \mathcal{F}_{\text{sel}} \cup \{f^*\}$, $\mathcal{F}_{\text{cand}} \leftarrow \mathcal{F}_{\text{cand}} \setminus \{f^*\}$
  **end**
  $\mathcal{F}^{\text{opt}}_{\text{QCI}X} \leftarrow \mathcal{F}_{\text{sel}}$
  $\theta_{\text{QCI}X} \leftarrow$ BNN trained on $\mathcal{F}^{\text{opt}}_{\text{QCI}X}$ using $RMSE$ (Eq.(5))
**end**
**for** *each new observation* $i$ **do**
  Predict latency: $\widehat{\overline{L_{\text{QCI}X}}}^{(i)} \leftarrow \theta_{\text{QCI}X}(\mathcal{F}^{\text{opt},(i)}_{\text{QCI}X})$
  Estimate exponential rate: $k^{(i)} \leftarrow 1/\widehat{\overline{L_{\text{QCI}X}}}^{(i)}$
  Compute LR:
  $\widehat{LR_{\text{QCI}X}}^{(i)} \leftarrow 1 - e^{-k^{(i)}(L^{(\max)}_{\text{QCI}X} - T_s)}$ (Eq.(7))
**end**

During exploitation stage, estimating LR in a new cell or ROP only requires inferring the BNN for a single datapoint, deriving the latency PDF and computing LR. The worst-time complexity of BNN inference with $S$ Monte-Carlo iterations is $\mathcal{O}\big(S \sum_{l=1}^{N_l} N_n(l)\, N_n(l-1)\big)$, where $l$ denotes layer index. The worst-time complexity of evaluating an exponential PDF is $\mathcal{O}(1)$ (i.e., constant time). These tasks can be executed fast per gNB. Finally, the BNNs must be retrained only when an event changing the relationship between predictors and the target variable occurs (e.g., update of packet scheduling policy, launch of new service...). Such sporadic events take place on a monthly or yearly basis.

## VI. Performance Assessment

In this section, the performance of the presented predictive LR framework is assessed over a dataset collected in a live cellular network. For clarity, the dataset is first introduced, experimental setup is next described, results are discussed later and, finally, execution times are outlined.

### A. Dataset description

Since comprehensive 5G data from large and congested networks covering diverse scenarios is not available yet, the dataset used in this work was gathered in a commercial LTE network operated by an Italian MNO. Specifically, cell-level CMs and PMs were collected every 15 minutes for two days in January 2023 from approximately 2000 cells covering the province of Bologna. Those cells operate on two frequency layers, hereafter referred to as layers T and M. Cells in layer T

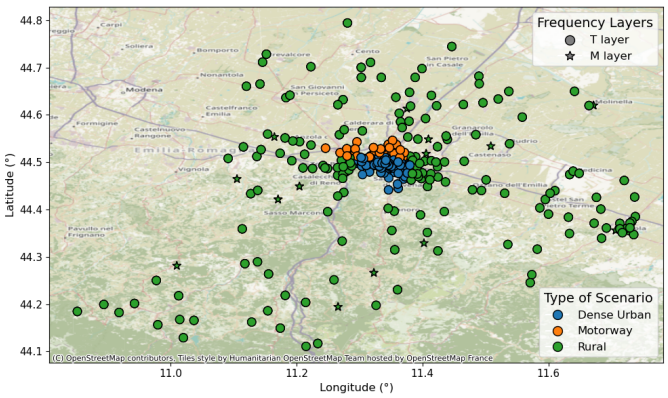



Fig. 3: Location of cell towers used for KPIs collection.

(1200 cells) operate at 1800 MHz with a 20-MHz bandwidth, whereas cells at layer M (800 cells) operate at 2100 MHz with a 15-MHz bandwidth. As illustrated in Fig. 3, cell towers are deployed over an area of approximately 4500 km$^2$. Thus, the dataset encompasses highly heterogeneous measurements from various scenarios, including dense urban, urban, rural and motorway environments.

After removing datapoints from cells with insufficient activity (cells with very few active UEs for QCI1 and QCI7), the dataset comprised $256 \times 10^4$ datapoints, each including time-stamp, cell layer, cell identifier and the set of 25 features listed in Table I. Nonetheless, it was observed that, for both layer T and M, many cells showed low *prb_util* values (i.e., underutilized cells). Note that an extremely imbalanced dataset can lead to sub-optimal performance of SL models. Moreover, MNOs show great interest for latency prediction in congested cells, prone to SLA violations. A correlation analysis presented later reveals that many KPIs in Table I show a higher correlation with $\overline{L_{\text{QCI}X}}$ when *prb_util* is high, suggesting that samples with high *prb_util* provide extremely valuable insights for model training. Consequently, random samples with low *prb_util* were filtered out to ensure a balanced dataset. The resulting dataset comprises $70 \times 10^4$ samples ($48 \times 10^4$ from layer T and $22 \times 10^4$ from layer M). 80% of those datapoints are used for model training and the remaining 20% are used for model testing.

### B. Experimental setup

The framework is implemented in Python, leveraging libraries specifically suited for ML and statistical analysis. SL models are implemented with *TensorFlow* and *TensorFlow Probability*, while *scikit-learn* is used for data processing. Finally, *shap* library is used for model explainability. Assessment procedure is next detailed.

**1) Correlation analysis.** As an initial step, a correlation analysis is conducted on the 22 input features listed in Table I, identified as the most closely linked to the critical mechanisms that exacerbate packet delays in the RAN. The purpose is twofold: (i) to evaluate the actual relevance of the selected features to compute mean packet latency, and (ii) to assess whether the impact of these features on mean packet latency varies across service classes, QCI1 and QCI7. To achieve this goal, the Spearman's rank correlation coefficient between

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



each input feature and $\overline{L_{\text{QCI}X}}$ was computed. Spearman's rank correlation is a non-parametric measure that evaluates the strength and direction of a monotonic relationship between two variables, capturing both linear and non-linear relationships among data [55]. As will be shown later, correlations observed for QCI1 and QCI7 differ significantly. Consequently, it was deemed appropriate to infer latency separately for each service class.

**2) SL-based mean latency estimation.** Then, the performance of the BNN described in Section V-B to estimate $\overline{L_{\text{QCI}X}}$ is assessed. Models considering the complete input feature set (i.e., 22 KPIs) are hereafter referred to as "full", while those considering only the optimal reduced feature set provided by SFS are referred to as "reduced". Two benchmark SL models are considered. The first is eXtreme Gradient Boosting (XGBoost), previously applied in [34] for throughput estimation using cell-level KPIs. The second is the BPNN used in [16] for latency estimation, consisting of dense variational layers with a configurable number of neurons, followed by a dense layer with 1 neuron that produces the rate parameter, which is then fed into an exponential layer to output $\widehat{f_{\overline{L_{\text{QCI}X}}}(t)}$ per cell. For these approaches, only the full model per QCI is assessed since, as will be shown later, they are outperformed by the framework proposed here. Hence, a total of 8 models are analyzed (2 QCIs · (1 XGBoost full + 1 BPNN full + 1 BNN full + 1 BNN reduced)).

Table II breaks down ANN architecture and main hyperparameter configuration for full models trained with BNN and BPNN for each QCI. Both BPNNs employ *Leaky Rectified Linear Unit (Leaky ReLU)* activation function in hidden layers, whereas the dense layer applies a *SoftPlus* activation function to ensure non-negative values when estimating the rate of the exponential distribution. Likewise, BNNs use a *Sigmoid* and *Rectified Linear Unit (ReLU)* activation functions in hidden and output layers, respectively. For both approaches, the optimal number of hidden layers, number of neurons per layer, and learning rate vary per QCI. Models are trained using the Adaptive Moment Estimation (ADAM) optimizer with a batch size of 128 datapoints. The maximum number of training epochs is set to $30 \times 10^4$. An early stopping condition is set to avoid overfitting. As a probabilistic model, the loss function used to train BPNNs is NLL, expressed as $\text{NLL} = -\sum_{i=1}^{N} \log p(y_i|\theta(x_i))$, with $p(y_i|\theta(x_i))$ the predicted probability of $y_i$ given the model parameters [54]. In contrast, as commented above, $RMSE$ serves as loss function to train both XGBoost and BNNs.

The best regression model per QCI is selected as a trade-off between model accuracy, complexity and explainability. Model accuracy is assessed using two key metrics. The first is the $MAE$ defined in (6), providing a direct interpretation of mean prediction error. The second is the trimmed Mean Absolute Percentage Error, $tMAPE$, defined as

$$tMAPE = \frac{100}{n_t} \sum_{i \in \mathcal{I}_t} \left| \frac{y_i - \hat{y}_i}{y_i} \right| , \tag{8}$$

where $\mathcal{I}_t$ represents the set of indices corresponding to the trimmed dataset and $n_t$ is the number of datapoints remaining after trimming. $tMAPE$ is computed by eliminating 10% of the highest and lowest errors, thus offering a robust measure of accuracy under typical network conditions. This is particularly relevant in the context of this work, where occasional extreme delays, caused by factors such as network congestion or high mobility, can disproportionately impact figures of merit.

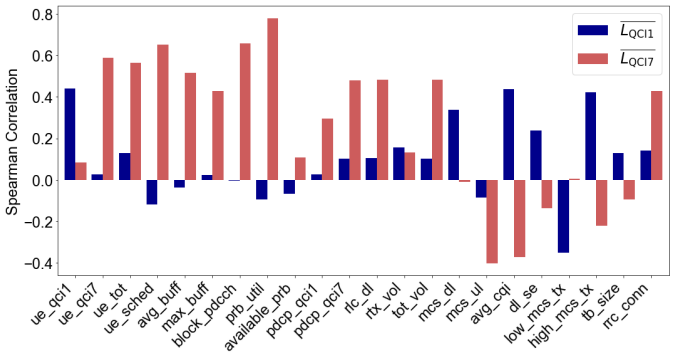



Fig. 4: Correlation among candidate features and $\overline{L_{\text{QCI}X}}$.

Model complexity is measured by the training time and the number of input features, as a proxy to (i) the backhaul load caused by data transfer between gNBs and OSS, (ii) the data storage capacity required in the OSS, and (iii) Central Processing Unit (CPU) capacity required for model training. Finally, model explainability is assessed using Shapley Additive Explanations (SHAP) values. SHAP is a model-agnostic method from Game Theory that quantifies the contribution of each feature to the model outcome. For this purpose, it assigns an importance score to each feature, indicating whether it increases or decreases the predicted value of the target variable and by how much [56].

**3) Latency-reliability evaluation.** Finally, LR for QCI1 is computed for each cell and ROP. To this end, for each datapoint, $\widehat{f_{\overline{L_{\text{QCI1}}}}(t)}$ is derived from $\overline{L_{QCI1}}$, and $\widehat{LR_{\text{QCI}X}}$ is then inferred using (7) with the latency threshold $L_{\text{QCI}X}^{(max)}$ set according to the requirement of the target service class, i.e., 100 ms for VoLTE. It should be pointed out that the accuracy of LR estimates can only be checked for QCI1, since LTE vendor equipment only store a LR KPI for VoLTE traffic.

### C. Results

Fig. 4 shows the mean Spearman's correlation between the 22 KPIs selected as candidate input features and mean latency KPIs, $\overline{L_{\text{QCI1}}}$ (blue bars) and $\overline{L_{\text{QCI7}}}$ (red bars). The fact that most KPIs present relevant correlation values (i.e., higher than 0.25 in absolute terms) with the target variable for some (or both) QCIs confirms the relevance of the selected features for latency estimation, validating their inclusion in the predictive framework. The only exceptions are *available_prb* and *tb_size*. The former is especially remarkable, since available cell bandwidth is well-related to cell throughput, and is thus often considered in previous cell performance models for user/cell throughput (e.g., [30] [34]). This difference in the dependencies of KPIs on cell state stresses the need for data-driven models specific for latency. It is also remarkable that many correlation values in Fig. 4 differ significantly per QCI. VoLTE traffic (QCI1), with its stringent latency

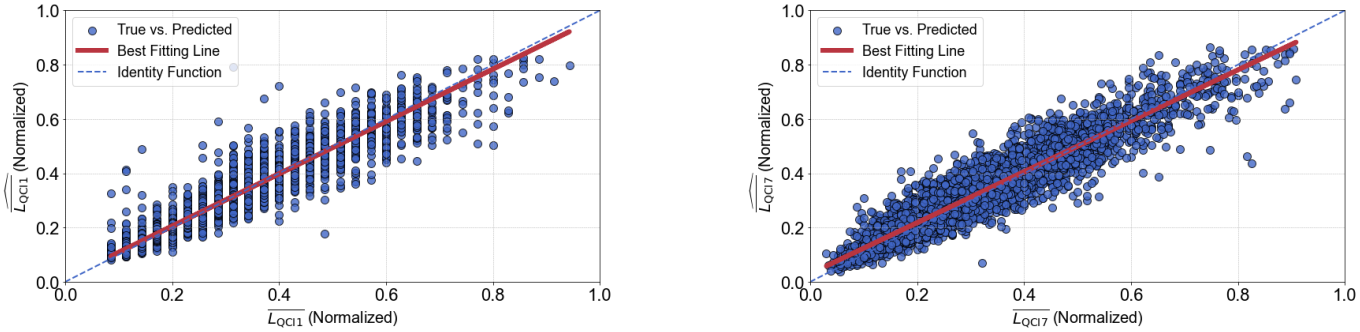
Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



TABLE II: Architecture and hyperparameters of BPNN and BNN models for estimating mean latency.

| Model | Architecture | Activation functions | Optimizer |
|---|---|---|---|
| BPNN QCI1 | 2 Dense Variational (128, 64) + 1 Dense (1) + 1 Exponential | *LeakyReLU* (hidden), *SoftPlus* (output) | Adam (LR=0.0001) |
| BPNN QCI7 | 3 Dense Variational (128, 64, 32) + 1 Dense (1) + 1 Exponential | *LeakyReLU* (hidden), *SoftPlus* (output) | Adam (LR=0.0001) |
| BNN QCI1 | 3 Dense Variational (128, 64, 32) + 1 Dense (1) | *Sigmoid* (hidden), *ReLU* (output) | Adam (LR=0.001) |
| BNN QCI7 | 2 Dense Variational (128, 64) + 1 Dense (1) | *Sigmoid* (hidden), *ReLU* (output) | Adam (LR=0.001) |




(a) QCI1. (b) QCI7.

Fig. 5: Dispersion diagram of predicted vs. true mean latency per QCI.

TABLE III: Performance of SL models to predict $\overline{L_{QCIX}}$.

| SL model | $MAE$ | $tMAPE$ | No. input features |
|---|---|---|---|
| XGBoost QCI1 Full | 1.38 ms | 10.98 % | 22 |
| XGBoost QCI7 Full | 9.98 ms | 13.88 % | 22 |
| BPNN QCI1 Full | 1.37 ms | 11.96 % | 22 |
| BPNN QCI7 Full | 8.22 ms | 14.21 % | 22 |
| BNN QCI1 Full | 0.66 ms | 6.63 % | 22 |
| BNN QCI7 Full | 8.19 ms | 10.75 % | 22 |
| BNN QCI1 Reduced | 0.67 ms | 6.76 % | 6 |
| BNN QCI7 Reduced | 8.08 ms | 10.93 % | 13 |

requirements, shows weaker correlations with traffic-related features compared to QCI7, suggesting that QCI1 prioritization mechanisms mitigate the effects of traffic load on packet delivery. Conversely, QCI7 latency shows stronger correlations with most features, revealing greater sensitivity to both channel and traffic variations. The strong correlation between *block_pdcch* and $\overline{L_{QCI7}}$ (i.e., 0.66), compared to the very low correlation of that feature with $\overline{L_{QCI1}}$ (i.e, 0.003), further illustrates this difference. In extreme cases, not only the value but also the sign of correlation differs per QCI. For instance, the opposite correlation trend for *avg_cqi*, *dl_se* and *high_mcs_dl* (positive for $\overline{L_{QCI1}}$ and negative for $\overline{L_{QCI7}}$) suggest that, in cells with good radio channel conditions, the QoS/channel-aware packet scheduler used in the network may be prioritizing traffic from QCI7, with very high potential throughput. This turns into an increase of latency for QCI1 packets, whose performance should not be compromised thanks to the high spectral efficiency. In contrast, when *low_mcs_tx* is high (i.e., poor radio channel conditions), QCI1 traffic is prioritized due to its stringent latency needs, resulting in higher latency for non-GBR traffic like QCI7, which can tolerate delays without strict bit rate guarantees.

To select the best combination of SL algorithm and input feature set for mean packet latency estimation, Table III breaks down accuracy and complexity metrics obtained when predicting $\overline{L_{QCI1}}$ and $\overline{L_{QCI7}}$ with full XGBoost, full BPNN, full BNN, and reduced BNN. Full models are first compared. BNNs outperform BPNNs and XGBoost in terms of $MAE$ and $tMAPE$ across both QCIs. Specifically, $MAE$ decreases from 1.38 ms (XGBoost) and 1.37 ms (BPNN) to 0.66 ms (BNN) for QCI1, and from 9.98 ms (XGBoost) and 8.22 ms (BPNN) to 8.19 ms (BNN) for QCI7. Likewise, $tMAPE$ decreases from 10.98% (XGBoost) and 11.96% (BPNN) to 6.63% for QCI1, and from 13.88% (XGBoost) and 14.21% (BPNN) to 10.75% for QCI7. These results confirm the suitability of BNN over the two baselines (BPNN and XGBoost) for the targeted problem.

Furthermore, the SFS process reduces the input feature set for the BNN approach. In Table I, ticks (✓) under the "SFS QCI1" and "SFS QCI7" columns indicate which input features are used in the reduced models for predicting $\overline{L_{QCI1}}$ and $\overline{L_{QCI7}}$, respectively. It can be observed that the subset of KPIs retained after SFS differ per QCI in both number (6 and 13 predictors for QCI1 and QCI7, respectively) and nature. In fact, only traffic-related KPIs (*ue_tot*, *prb_util*, *rlc_dl*, and *rrc_conn*) are used in both cases. As anticipated by the correlation analysis, QCI7 latency is more influenced by KPIs related to radio conditions, such as *avg_cqi*, *mcs_dl*, and *low_mcs_tx* (and hence the need for a larger set of predictors). Conversely, although these KPIs also show high correlation with $\overline{L_{QCI1}}$, they are not selected in the QCI1 model. The reason is that SFS evaluates features based on their joint predictive value rather than individual correlation strength. Thus, the selected features are not necessarily those with the highest correlation scores.

For both QCIs, the full and reduced BNNs deliver nearly identical performance (i.e., $MAE$ of 8.19 ms (full) vs. 8.08 ms (reduced) for QCI7, and $MAE$ of 0.66 ms (full) vs. 0.67 ms (reduced) for QCI1), while drastically reducing the complexity of model training and minimizing the operational burden on

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



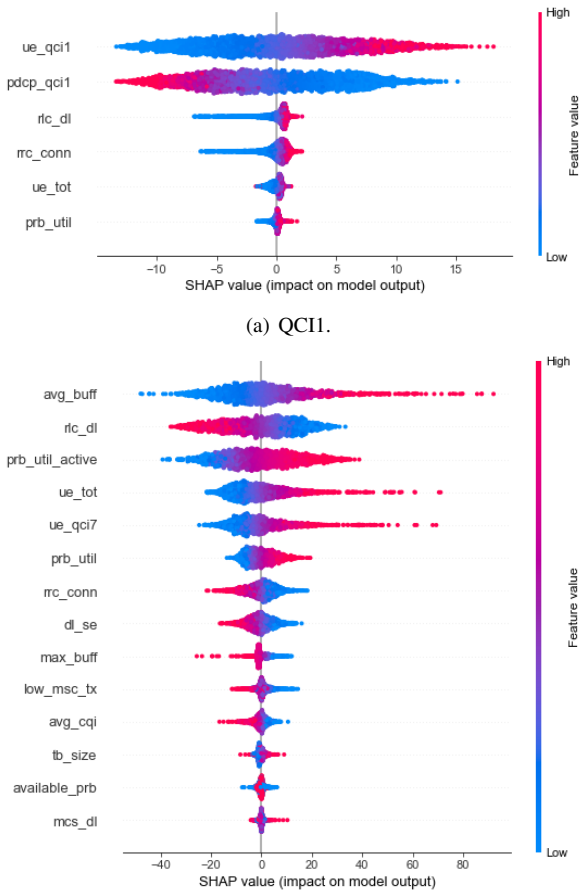
cellular networks. It is noteworthy that $MAE$ and $tMAPE$ values obtained for QCI7 are naturally higher due to the greater latency values experienced by users demanding non-GBR services. Nonetheless, accuracy metrics obtained with the best-performing models (BNNs with reduced feature set) reveal adequate model performance, with $MAE$=8.08 ms and $tMAPE$=10.93% for QCI7, and $MAE$=0.67 ms and $tMAPE$=6.76% for QCI1.

For a deeper analysis, Fig. 5.a) and 5.b) depict the real vs. estimated $\overline{L_{\text{QCI1}}}$ and $\overline{L_{\text{QCI7}}}$ values obtained with BNN reduced models. For confidentiality reasons, latency values are normalized. In both cases, the best-fitting line (red line) closely aligns with the identity function, showing a highly accurate match between predicted and actual latency values.

For model explainability, Fig. 6.a) and 6.b) present *Beeswarm* summary plots of Shapley analysis for reduced BNN models estimating $\overline{L_{\text{QCI1}}}$ and $\overline{L_{\text{QCI7}}}$, respectively. In these diagrams, features are ranked by their impact on model output. For each predictor, Shapley values of 50% of the test set are shown on the x-axis with round markers, colored by the normalized feature value. Specifically, blue means 0 (lowest), while red means 1 (highest).

Fig. 6.a) shows that the most influential KPIs for $\overline{L_{\text{QCI1}}}$ are the mean number of active users with QCI1 traffic, *ue_qci1*, and the mean number of QCI1 PDCP SDUs transmitted in DL, *pdcp_qci1*. As expected, an increase in the first predictor is associated with high $\widehat{\overline{L_{\text{QCI1}}}}$ values. However, an unexpected behavior is observed for *pdcp_qci1*, whose increase leads to low $\widehat{\overline{L_{\text{QCI1}}}}$ values. This anomaly can be explained by the strong correlation (0.61) between *ue_qci1* and *pdcp_qci1* KPIs. As pointed out in [57], highly correlated predictors can lead to misleading SHAP values, affecting not only the importance but also the direction of feature impact. This is due to the assumption of predictor independence in most approximation methods of SHAP values (e.g., Shapley sampling values or Kernel SHAP) for computational efficiency [58]. In this case, *rlc_dl* preserves an expected behavior, indicating that as *rlc_dl* decreases, latency decreases.

Fig. 6.b) exhibits a more extensive set of impactful features for $\overline{L_{\text{QCI7}}}$. Leading the list is *avg_buff*, indicating that RLC buffer status is the primary driver of high $\widehat{\overline{L_{\text{QCI7}}}}$ values. This perfectly aligns with the insights gained from the full-stack analysis in Section III and the mathematical characterization in Section IV-C, both of which identified the RLC layer and its sojourn time as the main bottleneck for RAN latency. Monitoring *avg_buff* KPI is thus critical for MNOs, as it not only reflects network congestion but may also highlight issues related to buffer dimensioning, potentially revealing instances of bufferbloat, as discussed in Section III. Following, *rlc_dl* KPI exhibits the same unexpected behavior discussed earlier for *pdcp_qci1* in Fig. 6.a). In this case, high values of traffic-related KPI *rlc_dl* correspond to low $\widehat{\overline{L_{\text{QCI7}}}}$ values. This is another case of multicollinearity, as the same effect is observed for the number of Radio Resource Control (RRC) connections, expressed by the *rrc_conn* KPI, which has a correlation value of 0.81 with *rlc_dl*. The extremely high correlation between these two predictors leads to misleading trends. Nonetheless, for other traffic-related KPIs, which prove to be essential in the predictive model, trends are as expected and reasonable. Specifically, extensive PRB usage during active periods, *prb_util_active*, strongly impacts $\overline{L_{\text{QCI7}}}$, as does the number of active users in the cell, *ue_tot*, and those utilizing QCI7 services, *ue_qci7*. This indicates that QCI7 services suffer under congestion, as highlighted in the above correlation analysis. The less evident but still relevant impact of channel quality KPIs, such as *avg_cqi* and *low_mcs_dl*, on $\overline{L_{\text{QCI7}}}$ confirms how the QCI7 service class is i) more sensitive to channel quality variations, and ii) prioritized when channel conditions are favorable, in order to maximize throughput, as previously discussed in the correlation analysis. In fact, for both KPIs, their higher values lead to low $\widehat{\overline{L_{\text{QCI7}}}}$ values.




(a) QCI1.

(b) QCI7.

Fig. 6: SHAP analysis on BNN reduced models.

Finally, to assess the consistency of the proposed methodology for computing LR, Fig. 7 shows the relationship between ground-truth $\overline{L_{\text{QCI1}}}$ measurements with LR values ($LR_{\text{QCI1}}$, represented by blue curve) and estimated LR values ($\widehat{LR}_{\text{QCI1}}$, represented by red curve). In both cases, the curves represent the second-order regression lines derived from the results of all test datapoints. It can be observed that both curves show a similar decaying trend, verifying the power of the proposed methodology for LR estimation in cellular RANs. It is also remarkable that the red curve shows a constant shift upward along the Y-axis compared to the blue curve curve, i.e., LR

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



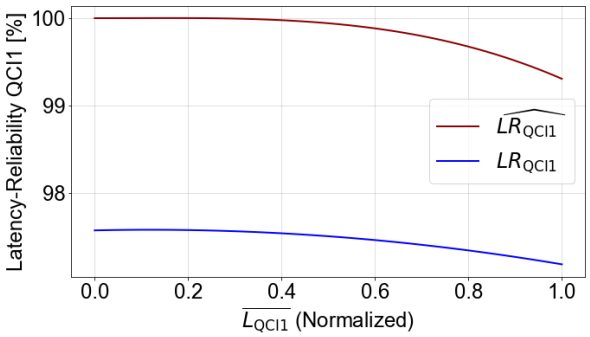



Fig. 7: $\widehat{LR_{\text{QCI1}}}$ from BNN reduced model vs. $LR_{\text{QCI1}}$ KPI.

estimates exceed actual measurements by 2.5% in relative terms for all $\overline{L_{\text{QCI1}}}$ values. This shift arises from the inclusion of delay introduced by HARQ retransmissions in ground-truth LR measurements. Such a procedure is excluded in the proposed framework since URLLC services (i.e., those where LR estimation is critical) are conceived to implement limited retransmission capabilities. Finally, note that LR results are only shown for QCI1, as LR KPI for QCI7 is not currently reported by vendor equipment. This limitation further motivates the need for a predictive framework like the one proposed here, which can infer LR even when it is not directly measurable.

### D. Execution times

Table IV summarizes training and inference times obtained for XGBoost, BPNN and BNN full models in this work, using a personal computer with an Intel Core i7-8700 processor working at 3.2 GHz with 16 GB of RAM. Hyperparameter optimization time is included. By comparing models for the same QCI, it can be observed that the BNN approach proposed here reduces training time by almost 10 times compared to both BPNN and XGBoost. These results confirm the improvement in model efficiency achieved by BNN.

Focusing on BNN results, it is noticeable that the full model predicting mean latency for QCI7 converges before that predicting mean latency for QCI1 (175 s vs. 452 s). Not shown in the table is the fact that, in both cases, models converge before reaching the maximum limit of $30 \times 10^4$ epochs, enabled by the early stopping mechanism (i.e., stop training when validation loss does not improve for 15 epochs). Specifically, validation loss becomes stable after 273 and 139 epochs for QCI7 and QCI1, respectively. These results confirm the convergence and stability of the proposed BNN models. Note that training time obtained for the worst BNN (i.e., 452 seconds) is adequate for the considered use case, where model training is performed offline. Under strict time constraints, model training can be speeded up by parallelization during hyperparameter optimization. Nonetheless, not shown in Table IV is the fact that training time decreases up to 35% with reduced models. Regarding inference time, in this work, total LR prediction time per datapoint is lower than 1 second for the best approach (BNN reduced models). Such a time meets the requirements of most RAN optimization use cases.

TABLE IV: Training and inference time of full SL models.

| SL model | Training time [s] | Inference time [s] |
|---|---|---|
| XGBoost QCI1 Full | 2525.16 | $9 \cdot 10^{-6}$ |
| XGBoost QCI7 Full | 1959.85 | $3 \cdot 10^{-6}$ |
| BPNN QCI1 Full | 3273.00 | 4.65 |
| BPNN QCI7 Full | 1194.22 | 3.25 |
| BNN QCI1 Full | 452.56 | 0.85 |
| BNN QCI7 Full | 175.03 | 0.34 |

Additional tests on a server-grade environment showed an average 40% reduction in execution times across all models, while preserving the same performance trends reported here.

## VII. Conclusions

As 6G draws closer, the demand for ultra-low latency calls for advanced mitigation strategies targeting key latency components. This work has presented a novel hybrid framework that combines analytical modeling and ML to estimate LR in the RAN using cell-level KPIs collected from 5G vendor equipment on a ROP, cell and QCI/5QI basis. For this purpose, a preliminary full-stack analysis of gNB operations and a queuing-theoretic latency model have been conducted to identify critical delay stages and map them to relevant KPIs among those currently available in vendor equipment. These steps have enabled a systematic mapping between delay-inducing mechanisms within the gNB and measurable vendor KPIs, allowing for targeted interventions and relieving MNOs from the operational burden of continuously monitoring large volumes of data. Then, the selected KPIs are used as input to a SL model to estimate mean latency. LR is subsequently derived leveraging an exponential distribution obtained by the above-mentioned analytical model of RAN latency.

Validation has been conducted over a real network dataset, considering latency experienced by users demanding services with QCI1 (VoLTE) and QCI7 (live video streaming, interactive gaming, etc.). A preliminary KPI correlation analysis has demonstrated the need for separate models per service class, due to differing predictor relevance. BNN has outperformed two baselines (XGBoost and BPNN) in the full model configuration. The best accuracy-computational efficiency trade-off is achieved by the reduced BNNs obtained via SFS, with $MAE$ and $tMAPE$ of 0.67 ms and 6.76% for QCI1 and 8.08 ms and 10.93% for QCI7. The set of KPIs selected as impactful predictors differs per QCI, pointing the need for a separate analysis per service class. Specifically, the 6 relevant predictors for QCI1 in this work include traffic and capacity KPIs, whereas the 13 relevant predictors for QCI7 comprise traffic, channel, and capacity KPIs. These mean latency estimates have allowed to accurately predict LR for QCI1 traffic in the absence of ground-truth measurements.

SHAP analysis has further confirmed the impact of different predictors used for the two distinct service classes, reinforcing the insights from the preliminary full-stack analysis and the mathematical characterization. Results have shown that, even if Explainable Artificial Intelligence (XAI) methods can be valuable tools for MNOs to explain model behavior by revealing dependencies and the influence of specific features on

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



latency, they must be applied with caution in the presence of highly correlated predictors. The limited number of features helps mitigate potential issues related to multicollinearity.

To support deployment in real-world environments, the framework has been designed to operate in the OSS, where data from all gNBs, required to train the SL-based LR estimation models, is available. However, with recent advances on federated learning, these SL models can be trained in a distributed manner (e.g., by combining several models trained in data centres) [59]. Such an approach reduces the volume of data transferred across the network, thereby alleviating congestion in the backhaul and improving scalability. Additionally, federated learning can accelerate the overall training process by leveraging parallel computation in data centers. Once SL models are trained, the proposed system can be exploited in a distributed fashion, ideally on a per gNB basis.

Future work will explore feature extraction methods (e.g., principal component analysis, autoencoder...) to derive a minimal set of uncorrelated predictors to reduce training times and required dataset size. To enhance interpretability, more elaborated SHAP analysis methods that do not assume feature independence (e.g., extended Kernel SHAP, SHAP cohort refinement...) will be used.

## ACKNOWLEDGMENTS

This work was funded by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, through the partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"), and the Spanish Ministry of Science and Innovation (PID2021-122217OBI00/AEI/10.13039/501100011033).

## REFERENCES

[1] Ericsson, "Ericsson Mobility Report: Business Review 2024." Tech. Rep.

[2] S. Kumar, "AI/ML Enabled Automation System for Software Defined Disaggregated Open Radio Access Networks: Transforming Telecommunication Business," *Big Data Mining and Analytics*, vol. 7, no. 2, pp. 271–293, 2024.

[3] F. Wu, F. Lyu, H. Wu, J. Ren, Y. Zhang, and X. Shen, "Characterizing User Association Patterns for Optimizing Small-Cell Edge System Performance," *IEEE Network*, vol. 37, no. 3, pp. 210–217, 2023.

[4] M. Banafaa *et al.*, "6G Mobile Communication Technology: Requirements, Targets, Applications, Challenges, Advantages, and Opportunities," *Alexandria Engineering Journal*, vol. 64, pp. 245–274, 2023.

[5] A. A. Shamsabadi *et al.*, "Exploring the 6G Potentials: Immersive, Hyper Reliable, and Low-Latency Communication," Tech. Rep., 2024.

[6] B. Hassan, S. Baig, and M. Asif, "Key Technologies for Ultra-Reliable and Low-Latency Communication in 6G," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 106–113, 2021.

[7] I. Parvez *et al.*, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[8] U. Bauknecht and T. Enderle, "An Investigation on Core Network Latency," in *2020 30th International Telecommunication Networks and Applications Conference (ITNAC)*, 2020, pp. 1–6.

[9] J. Qiao *et al.*, "Research on Diagnosis System of 5G Data Service Latency Problem," in *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2023, pp. 2181–2186.

[10] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies (Rel. 18)," Tech. Rep. TR 38.913 V18.0.0, 2024.

[11] D. Maaz, A. Galindo-Serrano, and S. E. Elayoubi, "URLLC User Plane Latency Performance in New Radio," in *2018 25th International Conference on Telecommunications (ICT)*, 2018, pp. 225–229.

[12] M. Indoonundon and T. P. Fowdur, "Latency Components and Analysis in 5G New Radio," in *2022 4th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, 2022, pp. 1–8.

[13] Y. Zhao, M. Wei, C. Hu, and W. Xie, "Latency Analysis and Field Trial for 5G NR," in *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2022, pp. 1–5.

[14] M. El Rajab, L. Yang, and A. Shami, "Zero-Touch Networks: Towards Next-Generation Network Automation," *Computer Networks*, vol. 243, p. 110294, 2024.

[15] M. Stojčić *et al.*, "Predictive Modeling of Delay in an LTE Network by Optimizing the Number of Predictors Using Dimensionality Reduction Techniques," *Applied Sciences*, vol. 13, no. 14, 2023.

[16] M. Skocaj *et al.*, "Data-driven Predictive Latency for 5G: A Theoretical and Experimental Analysis Using Network Measurements," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–6.

[17] S. Mostafavi, G. P. Sharma, and J. Gross, "Data-Driven Latency Probability Prediction for Wireless Networks: Focusing on Tail Probabilities," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 4338–4344.

[18] H. N. Qureshi *et al.*, "Service Level Agreements for 5G and Beyond: Overview, Challenges and Enablers of 5G-Healthcare Systems," *IEEE Access*, vol. 9, pp. 1044–1061, 2021.

[19] 3GPP, "Technical Specification Group Services and System Aspects; Service Requirements for the 5G System; Stage 1 (Rel. 20)," Tech. Rep. TR 22.261 V20.0.0, 2024.

[20] X. Jiang *et al.*, "Low-Latency Networking: Where Latency Lurks and How to Tame It," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 280–306, 2019.

[21] A. Carvalho, L. M. Correia, A. Grilo, and R. Dinis, "Analysis of Strategies for Minimising End-to-End Latency in 5G Networks," in *2022 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom)*, 2022, pp. 1–6.

[22] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G New Radio Numerologies and their Impact on the End-To-End Latency," in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2018, pp. 1–6.

[23] M. Mhedhbi, S. Elayoubi, and G. Leconte, "AI-Based Prediction for Ultra Reliable Low Latency Service Performance in Industrial Environments," in *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2022, pp. 130–135.

[24] N. Zhang, P. He, Z. Wu, P. Chen, L. Wang, and Z. Ye, "Latency Analysis and Trial for 5G Ultra Reliable Low Latency Communication," in *2023 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, 2023, pp. 1–6.

[25] M. Seidel *et al.*, "On the Impact of 5G User Equipments on Latency across Chipset Generations," in *2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2024, pp. 177–185.

[26] 5G Americas, "New Services & Applications with 5G Ultra-Reliable Low Latency Communications," Tech. Rep., 2018.

[27] R. A. K. Fezeu *et al.*, "An In-Depth Measurement Analysis of 5G mmWave PHY Latency and Its Impact on End-to-End Delay," in *Passive and Active Measurement: 24th International Conference, PAM 2023, Virtual Event, March 21–23, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023.

[28] M. Abdullah, S. E. Elayoubi, T. Chahed, and A. Lisser, "Performance Modeling and Dimensioning of Latency-Critical Traffic in 5G Networks," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 4307–4312.

[29] M. C. Lucas-Estañ *et al.*, "An Analytical Latency Model and Evaluation of the Capacity of 5G NR to Support V2X Services Using V2N2V Communications," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2293–2306, 2023.

[30] V. Wille, M. Toril, and S. Luna-Ramirez, "Estimating Pole Capacity in a Live HSDPA Network," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1260–1263, 2013.

[31] D. Parracho, D. Duarte, I. Pinto, and P. Vieira, "An Improved Capacity Model Based on Radio Measurements for a 4G and Beyond Wireless Network," in *2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2018, pp. 314–318.

[32] T. ur Rehman, M. A. I. Baig, and A. Ahmad, "LTE Downlink Throughput Modeling Using Neural Networks," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2017, pp. 265–270.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3603923



[33] C. Gijón *et al.*, "Estimating Pole Capacity from Radio Network Performance Statistics by Supervised Learning," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2020.
[34] C. Gijón, M. Toril, and S. Luna-Ramírez, "Data-Driven Estimation of Throughput Performance in Sliced Radio Access Networks via Supervised Learning," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1008–1023, 2023.
[35] D. Minovski, N. Ögren, K. Mitra, and C. Åhlund, "Throughput Prediction Using Machine Learning in LTE and 5G Networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1825–1840, 2023.
[36] F. S. D. Silva *et al.*, "Proactive ML-Assisted and Quality-Driven Slice Application Service Management to Keep QoE in 5G Mobile Networks," in *2023 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2023, pp. 182–184.
[37] 3GPP, "System Architecture for the 5G System (5GS); Stage 2," Tech. Rep. TS 23.501 V18.3.0, 2023.
[38] C. Baena, S. Fortes, E. Baena, and R. Barco, "Estimation of Video Streaming KQIs for Radio Access Negotiation in Network Slicing Scenarios," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1304–1307, 2020.
[39] Z. Xing, H. Li, W. Liu, Z. Ren, J. Chen, J. Xu, and C. Qin, "Spectrum Efficiency Prediction for Real-World 5G Networks Based on Drive Testing Data," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2136–2141.
[40] D. López-Pérez, A. De Domenico, N. Piovesan, and M. Debbah, "Data-Driven Energy Efficiency Modeling in Large-Scale Networks: An Expert Knowledge and ML-Based Approach," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 780–804, 2024.
[41] S. Wang, F. Wu, J. Gao, S. Duan, F. Lyu, H. Wu, Y. Zhang, and X. S. Shen, "Dynamic RRH-BBU Mapping for C-RAN: A Data-Driven Approach," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023, pp. 2656–2661.
[42] J. A. Fernández-Segovia *et al.*, "Estimating Cell Capacity From Network Measurements in a Multi-Service LTE System," *IEEE Communications Letters*, vol. 19, no. 3, pp. 431–434, 2015.
[43] A. S. Khatouni, F. Soro, and D. Giordano, "A Machine Learning Application for Latency Prediction in Operational 4G Networks," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, pp. 71–74.
[44] M. Irazabal *et al.*, "Preventing RLC Buffer Sojourn Delays in 5G," *IEEE Access*, vol. 9, pp. 39 466–39 488, 2021.
[45] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, 1st ed. USA: Academic Press, Inc., 2018.
[46] S. Monikandan, A.Sivasubramanian, and S. Babu, "A Review of MAC Scheduling Algorithms in LTE System," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, pp. 1056–1068, 2017.
[47] 3GPP, "NR; Medium Access Control (MAC) Protocol Specification (Release 18)," 3GPP TS 38.321 V18.5.0, Mar. 2025.
[48] ——, "Feasibility Study for Evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN) (Rel. 18)," Tech. Rep. TR 25.912 V18.0.0, 2024.
[49] P. Cochois *et al.*, *5G19 Performances Optimisation Guidelines*, 1st ed., NetEng WebNEI.
[50] 3GPP, "Telecommunication Management; Performance Management (PM); Concept and Requirements (Rel. 18)," Tech. Rep. TS 32.401 V18.0.0, 2024.
[51] L. Kleinrock, *Queueing Systems, Vol. I: Theory*, 1975.
[52] I. T. Union, "ITU-TD Study Group 2 Report on Teletraffic Engineering," International Telecommunication Union (ITU), Geneva, Switzerland, Tech. Rep. D-STG-SG02.16.1, 2001.
[53] 3GPP, "NR; Physical Channels and Modulation (Release 18)," 3GPP TS 38.211 V18.6.0, Mar. 2025.
[54] D. T. Chang, "Probabilistic Deep Learning with Probabilistic Neural Networks and Deep Probabilistic Models," *ArXiv*, vol. abs/2106.00120, 2021.
[55] G. W. Corder and D. I. Foreman, "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach," 2009.
[56] *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988.
[57] A. M. Salih, "Explainable Artificial Intelligence and Multicollinearity: A Mini Review of Current Approaches," *ArXiv*, vol. abs/2406.11524, 2024.
[58] K. Aas, M. Jullum, and A. Løland, "Explaining Individual Predictions when Features are Dependent: More Accurate Approximations to Shapley Values," *Artificial Intelligence*, vol. 298, p. 103502, 2021.
[59] J. M. Parra-Ullauri *et al.*, "Federated Analytics for 6G Networks: Applications, Challenges, and Opportunities," *IEEE Network*, vol. 38, no. 2, pp. 9–17, 2024.

**Francesca Conserva** holds a BSc and MSc in Telecommunications Engineering and a PhD in Electronics, Telecommunications, and Information Technologies Engineering from the University of Bologna. Her doctoral research focused on RRM in B5G networks, contributing to the design of mobility-aware RRM algorithms for UAV-aided vehicular networks and to the development of AI-based predictive framework for RAN optimization using network KPIs. She now works as a Postdoctoral Researcher at WiLab (CNIT), where she focuses on Network Digital Twin technologies.

**Carolina Gijón** received her B.Sc. degree in Telecommunication Systems Engineering and her M.Sc. Degree in Telecommunication Engineering from the University of Málaga, Spain, in 2016 and 2018, respectively. Since 2017, she is a research assistant in the Communications Engineering Department of the University of Málaga, where she received the Ph.D. degree in 2023. In 2021, she was visiting researcher at King's College London. Her research interests include zero-touch networks, machine learning and radio resource management.

**Matías Toril** received his M.S in Telecommunication Engineering and the Ph.D degrees from the University of Málaga, Spain, in 1995 and 2007 respectively. Since 1997, he is Lecturer in the Communications Engineering Department, University of Málaga, where he is currently Full Professor. He has co-authored more than 130 publications in leading conferences and journals and 8 patents owned by Nokia or Ericsson. His current research interests include self-organizing networks, radio resource management and data analytics.

**Davide Micheli** received the "Dottore Ingegnere" degree in electronics engineering from the Università Politecnica delle Marche, in 2001, and the master's degree in astronautic engineering and Ph.D. degree in aerospace engineering from the "Sapienza" University of Rome, Italy, in 2007 and 2011, respectively. In 2014, he obtains the national scientific qualification as a university professor in Aerospace Engineering. He is currently with the Department of Wireless Access Engineering, Telecom Italia, Rome, Italy. He is also the author of numerous scientific articles in international journals.

**Maurizio Fodrini** Graduated in Telecommunications Engineering, he joined Telecom Italia in 2001 and became part of FiberCop in July 2024. He is currently engaged within the "Technology Plans and Innovation Programs" department, focusing on university collaborations, national and EU-funded research projects, and initiatives on the evolution toward autonomous networks, where he also contributes as a 3GPP delegate. He has been involved in the development of patents, RFI/RFQ documentation and international conference participation.

**Roberto Verdone** is Full Professor in Telecommunications at the University of Bologna since 2001. He founded a research group (Radio Networks) working on i) Radio Resource Management for mobile systems, ii) MAC, routing and topology of wireless sensor networks, iii) architectures and technologies for the IoT. He is currently active in the field of the Industrial IoT, 5G and B5G systems using THz communications, and UAV-Aided Mobile Radio Networks. Since 2020 he is Director of the CNIT National Laboratory of Wireless Communications, WiLab. Since 2021 he is co-Director of the WiLab-Huawei Joint Innovation Center on "Intelligent IoT for 6G". He published about 200 research papers, on IEEE journals/conferences.

Authorized licensed use limited to: National Huaqiao University. Downloaded on November 19,2025 at 07:33:07 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.