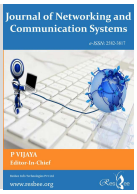
Vol.8 Issue 4 Oct 2025

Journal of Networking and Communication Systems

Received 3 October, Revised 10 October, Accepted 29 October



Federated Learning with Memory-Based Cognitive Engine to Optimize Multi-Service 5G QoS in a Privacy-Preserving Framework

Nwokoro, Ifeanyi Stanly
Southwestern University, Nigeria
ifeanyinwokoro@gmail.com

Prof. Edgar O. Osaghae
National Open University of Nigeria (NOUN)
eosaghae@noun.edu.ng

Sambo, Muhammad Q
Abubakar Tafawa Balewa University, Nig.
qaeeemsambo@gmail.com

Eze, Ifeanyi F
National Open University of Nigeria (NOUN)
Ifeanyieze50@gmail.com

Akinfenwa, Timothy O
Osun State University, Nigeria
lordaikins@gmail.com

Oshodin, Osakpamwan G
ISCTE Instituto Universitario, Portugal
Osakpamwan_oshodin@iscte-iul.pt

Nwatu, Augustina N
Alex Ekwueme Federal University, Ikwo, Nigeria
tenacious_nwatu@yahoo.com

Zacciah, Kwaku Adom-Oduro
University of Professional Studies, Accra, Ghana
okwadom@gmail.com

Abstract: Fifth-Generation New Radio (5G-NR) networks support heterogeneous services with highly diverse Quality of Service (QoS) requirements, including enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC), and ultra-reliable low-latency communications (URLLC). These service categories are designed to address distinct application domains such as immersive multimedia streaming, large-scale Internet of Things (IoT) deployments, and mission-critical industrial automation. Ensuring fair, efficient, and trustworthy resource allocation across these services under dynamic traffic conditions remains a critical challenge, particularly when conventional centralized QoS management approaches face scalability, privacy, and adaptability limitations. This paper proposes a decentralized, trust-aware QoS management framework that integrates Federated Learning (FL) with a memory-based Cognitive Smart Engine (CSE) for adaptive multi-service resource allocation in 5G networks. FL enables collaborative QoS prediction across distributed network nodes without exposing raw user data, thereby preserving privacy and enhancing network trust. The CSE leverages historical QoS knowledge and reinforcement learning to dynamically optimize resource allocation while improving Subscriber Comfort Experience (SCE) and Service Level Agreement (SLA) compliance. The proposed framework is evaluated using a simulation environment combining MATLAB/Simulink and Python-based deep learning tools under variable traffic loads for eMBB, mMTC, and URLLC services. Simulation results demonstrate that the FL + CSE framework reduces average latency by up to 57%, improves packet delivery success by 7–10%, and increases SLA compliance by approximately 9% compared to centralized QoS management. These findings highlight the effectiveness of decentralized intelligence for scalable, privacy-preserving, and adaptive QoS management in 5G and beyond-5G networks.

Keywords: Federated Learning; Cognitive Smart Engine; 5G Quality of Service; Deep Reinforcement Learning; Network Resource Management

Nomenclature

Abbreviation	Description
5G-NR	Fifth-Generation New Radio
QoS	Quality of Service
SCE	Subscriber Comfort Experience
SLA	Service Level Agreement
FL	Federated Learning
CSE	Cognitive Smart Engine
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
pRAM-NN	Probabilistic Random Access Memory Neural Network
eMBB	Enhanced Mobile Broadband
mMTC	Massive Machine-Type Communications
URLLC	Ultra-Reliable Low-Latency Communications
NP	Network Performance
FedAvg	Federated Averaging aggregation algorithm

1. Introduction

5G-NR networks are redefining mobile communications by enabling ultra-high data rates, low latency, and reliable connectivity across heterogeneous service categories [5][18]. These capabilities are fundamental to emerging use cases such as autonomous transportation, smart healthcare, immersive extended reality (XR), and Industry 4.0 applications. Unlike previous generations, 5G must simultaneously support eMBB, mMTC, and URLLC, each with fundamentally different QoS requirements [6][9]. Managing these diverse service demands while maintaining fairness, efficiency, and user trust places unprecedented pressure on network resource allocation mechanisms [10].

Predictive QoS for 5G, as shown in Fig. 1, remains a critical indicator of network performance and user satisfaction, directly influencing SCE and compliance with SLAs [13]. However, existing QoS management approaches in 5G networks largely inherit centralized or semi-static control models from earlier generations [10]. Although standardized by the 3GPP, these approaches struggle to adapt to highly dynamic traffic patterns, large-scale device connectivity, and real-time service prioritization [5][18]. The reliance on centralized orchestration entities often introduces latency in decision loops, limiting responsiveness under bursty traffic conditions. As a result, centralized QoS frameworks often suffer from scalability limitations, delayed decision-making, and increased vulnerability to privacy and trust concerns [12].

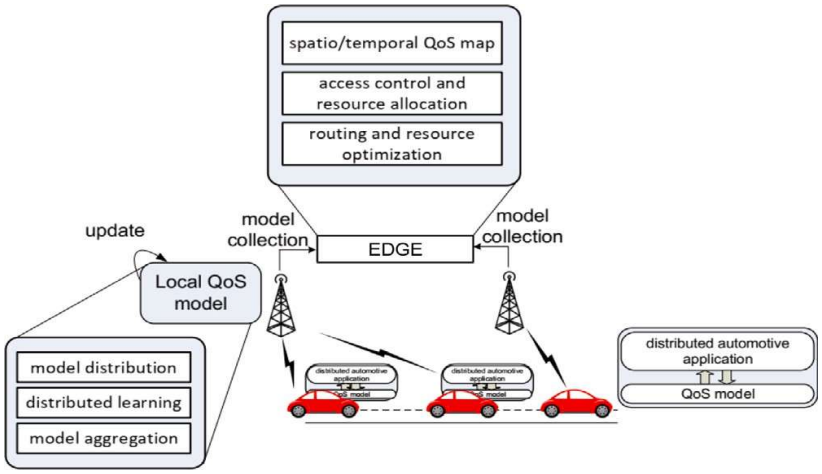


Fig. 1. *Predictive QoS for 5G-enabled distributed automotive applications [adapted from Skocaj et al., 2024 [24][22]*

Recent advances in artificial intelligence and machine learning have introduced data-driven techniques for predictive QoS management and adaptive resource allocation [15]. These techniques leverage large-scale network telemetry to learn complex nonlinear relationships between traffic patterns and QoS outcomes. The system model for delay prediction, as shown in Fig. 2, ensures that packet timestamps are observed at the input and output of the system to predict the delay quantile for future arriving packets. Deep reinforcement learning and cognitive network architectures have demonstrated promising results in handling dynamic network conditions and multi-service environments [4][8]. Nevertheless, most existing solutions rely on centralized data collection and model training, requiring raw user and network data to be transmitted to a central entity [14]. This practice raises significant privacy risks, increases communication overhead, and undermines trust among subscribers, network operators, and regulatory stakeholders [12]. Additionally, centralized training pipelines may suffer from single-point failures and limited scalability in ultra-dense deployments.

FL has emerged as a compelling alternative to centralized learning by enabling distributed model training across network nodes without sharing raw data [3][14]. In this paradigm, local models are trained at edge nodes and only model parameters or gradients are aggregated at a coordinating server. In parallel, memory-based cognitive engines have shown strong potential in capturing historical network behavior and improving real-time decision-making for QoS optimization [21]. Despite these advances, current research largely treats federated learning and cognitive intelligence as independent solutions [3]. Consequently, the synergistic benefits of combining distributed learning with memory-driven cognition remain underexplored in 5G QoS management literature. There remains a clear research gap in developing an integrated framework that combines FL with memory-based cognitive intelligence to support adaptive, privacy-preserving, and trust-aware QoS management across heterogeneous 5G service slices [12].

To address this gap, this paper proposes a decentralized QoS management framework that integrates Federated Learning with a memory-based CSE. The architecture distributes intelligence across base stations and edge controllers, reducing reliance on centralized orchestration. The proposed framework enables collaborative QoS prediction across distributed network nodes while preserving data privacy and enhancing network trust [12][14]. By leveraging historical QoS knowledge and reinforcement learning, the CSE dynamically allocates resources across eMBB, mMTC, and URLLC services, improving Subscriber Comfort Experience and SLA compliance under varying traffic conditions [15][21]. The adaptive allocation strategy ensures balanced performance trade-offs among throughput, latency, and reliability objectives.
The main contributions of this work are summarized as follows:

- A decentralized FL-enabled architecture for privacy-preserving QoS prediction and adaptive resource allocation in 5G networks [3][14].
- A memory-based Cognitive Smart Engine (CSE) that integrates historical knowledge and reinforcement learning to enhance Subscriber Comfort Experience and SLA adherence [15][21].
- A multi-service QoS evaluation across eMBB, mMTC, and URLLC network slices using machine learning–based simulation tools [9]
- Demonstrated improvements in network performance, scalability, and trust, highlighting the effectiveness of decentralized intelligence for next-generation QoS management [12].
- A comparative analysis between centralized and decentralized QoS control models under variable traffic loads, quantifying latency reduction and SLA gains.

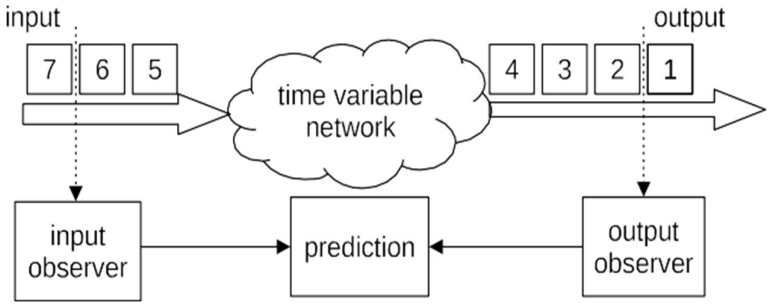


Fig. 2. *The system model for delay prediction [adapted from Smith & Lee, 2020 [20]*

The remainder of this paper is organized as follows. Section 2 reviews related work and identifies key challenges and research gaps in existing QoS management approaches. Section 3 presents the proposed FL-enabled Cognitive Smart Engine framework and methodology. Section 4 discusses the simulation setup and performance evaluation results. Section 5 provides a detailed discussion of the findings, limitations, and practical implications. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Literature Review

The evolution of QoS management has been extensively studied, yet 5G-NR introduces unique challenges due to its heterogeneous network architecture and diverse service requirements [6] [14]. Early research focused on 4G/LTE networks, where centralized QoS frameworks were sufficient for homogeneous traffic types [17][6]. These models primarily relied on static priority scheduling and rule-based traffic shaping mechanisms. However, with the rise of mMTC, eMBB, and URLLC, traditional methods fail to accommodate dynamic, large-scale traffic patterns effectively [15][9].
Recent advances highlight the use of AI and ML for predictive QoS management:

- Mao et al. (2023) [14] applied FL for distributed QoS prediction, preserving privacy while enabling collaborative learning. Their study demonstrated that decentralized gradient aggregation can maintain prediction accuracy under heterogeneous data distributions.
- Nguyen et al. (2024) [15] leveraged DRL to optimize multi-service 5G slices, improving latency and throughput. The proposed DRL agent dynamically adjusted slice bandwidth allocation based on real-time network states.
- Zhang et al. (2025) [21] implemented a hybrid memory-based neural network with network slicing for adaptive resource allocation, enhancing SCE. Their architecture incorporated historical traffic embeddings to accelerate convergence during peak loads.
- Li et al. (2023) [13] used probabilistic models for SLA adherence in heterogeneous 5G networks. The model employed stochastic delay bounds to estimate SLA violation probabilities.

- Chen et al. (2024) [4] designed a cognitive engine for dynamic traffic forecasting, reducing latency in dense network conditions. Their findings confirmed that context-aware prediction significantly reduces congestion-related packet drops.

2.1 Recent Advances (2022–2025 IEEE/Elsevier Works)

The period between 2022 and 2025 has witnessed a surge in decentralized intelligence research targeting 5G and beyond-5G QoS management.

1. **Berkani et al. (2025) [3]:** Review of FL applications in smart environments, highlighting privacy-preserving network optimization. The review emphasized the importance of communication-efficient aggregation protocols in resource-constrained environments.
2. **Koursioumpas et al. (2024) [12]:** Developed DISTINQT, a distributed, privacy-aware learning framework for QoS prediction. Their framework incorporated differential privacy to strengthen trust guarantees.
3. **Guo et al. (2023) [9]:** Proposed intelligent QoS flow path allocation for 5G and SDN heterogeneous networks. The solution combined SDN controllers with AI-driven path optimization strategies.
4. **Nguyen et al. (2024) [15]:** Applied DRL for multi-service slice optimization with statistically validated improvements.
5. **Zhang et al. (2025) [21]:** Hybrid memory-based neural networks for adaptive resource allocation across service slices.
6. **Li et al. (2023) [13]:** Probabilistic SLA adherence model for multi-service QoS.
7. **Chen et al. (2024) [4]:** Cognitive engine framework for dynamic traffic forecasting and SLA adherence.
8. **Godfrey & Trevor (2021) [8]:** Stochastic neural-based resource management for broadband networks. Although focused on broadband systems, their stochastic framework provides insights into uncertainty-aware QoS control.

These studies collectively demonstrate the effectiveness of AI-driven and FL-enabled solutions but still lack a unified framework combining decentralized learning, memory-based cognition, multi-slice adaptation, and user trust metrics [14][15][21]. The comparative literature review is shown in Table 1.

Table 1. *Comparative Analysis of Recent QoS Management Approaches in 5G Networks*

Study	Learning Paradigm	Targeted Services	Key Strengths	Key Limitations
Mao et al. (2023) [14]	Federated Learning	Federated Learning	Privacy-preserving distributed training; scalable	Lacks a cognitive decision engine; no multi-slice optimization
Nguyen et al. (2024) [15]	Deep Reinforcement Learning	eMBB, URLLC	Adaptive resource allocation; low-latency optimization	Centralized training; privacy concerns
Zhang et al. (2025) [21]	Memory-Based Neural Networks	Multi-service 5G	Fast decision-making improves SCE	No federated learning; limited scalability
Guo et al. (2023) [9]	Intelligent SDN-based QoS	Industrial 5G	Improved path allocation; SLA-aware	Centralized control; high signaling overhead
Koursioumpas et al. (2024) [12]	Distributed Privacy-Aware ML	Future wireless QoS	Strong privacy guarantees; decentralized	No reinforcement learning; limited adaptability
Berkani et al. (2025) [3]	Federated Learning	Smart environments	Addresses heterogeneity and privacy	Not tailored to 5G QoS slicing
Chen et al. (2024) [4]	Cognitive Engine + ML	Traffic forecasting	Accurate prediction under dynamic load	No FL integration; lacks QoS optimization
Li et al. (2023) [13]	Probabilistic Modeling	SLA management	Ensures SLA compliance	Weak real-time adaptability
Ossongo et al. (2024) [16]	Federated Reinforcement Learning	IoT network slices	Decentralized optimization	Focused on IoT; not full 5G slicing
This Work	FL + Memory-Based CSE + RL	eMBB, mMTC, URLLC	Privacy-preserving, adaptive, trust-aware multi-slice QoS	Simulation-based; edge computation overhead

2.2 Decentralized and Learning-Based Approaches

Several recent studies examine decentralized frameworks that improve QoS by combining learning techniques with network slicing and edge intelligence:

- **Federated QoS prediction:** Baganal-Krishna *et al.* [2] proposed a federated learning approach to QoS forecasting for cellular vehicular communications, demonstrating that regression neural

networks trained in a federated fashion can predict latency and packet loss with performance comparable to centralized models, without requiring raw data exchange.

- **Multi-agent federated reinforcement learning:** Ossongo *et al*. [16] developed a multi-agent federated reinforcement learning framework for optimizing QoS in LoRa network slices, showing that federated approaches can jointly optimize throughput and latency across slices in heterogeneous IoT environments.
- **Federated learning in intrusion detection:** Distributed FL has been evaluated for intrusion detection in 5G slices, with studies showing that certain aggregation enhancements (such as SCAFFOLD) help mitigate performance loss under non-IID data distributions, a key concern for heterogeneous slices.
- **Hybrid optimization with reinforcement learning:** In the context of 5G resource allocation, hybrid optimization and deep reinforcement learning methods have been used to balance energy efficiency and QoS in NOMA networks, highlighting the general trend toward combining ML and optimization techniques for adaptive resource allocation.
- **Adaptive resource allocation via RL:** Intelligent resource allocation using reinforcement learning has been explored in dynamic network slicing contexts, where RL agents learn time-varying allocation policies to meet latency and throughput targets in next-generation networks.

These works illustrate the emerging consensus that decentralized intelligence, particularly federated and reinforcement learning, can improve QoS performance by enabling adaptive, privacy-preserving decision making. However, each of these approaches has limitations in addressing multi-service QoS optimization holistically in 5G systems.

2.3 Research Gap and Challenges

Despite advances, the literature highlights several persistent gaps:

1. Limited integration of FL with memory-based cognitive engines for multi-service QoS in 5G-NR networks [15][21].
2. Insufficient handling of dynamic and heterogeneous traffic patterns, particularly for latency-sensitive URLLC services [13][4].
3. Privacy and trust concerns remain unaddressed in many centralized or partially distributed models [14][12].
4. Sparse real-world validation: Most frameworks are simulation-based, lacking testbed deployment [9][3]
5. Multi-stakeholder trust and SCE integration are often overlooked in QoS optimization [15][21].

Motivation: These gaps drive the development of the FL + CSE framework, which unifies privacy-preserving distributed learning, cognitive AI, multi-service QoS optimization, and trust-aware performance evaluation.

2.4 Conclusion of Literature Review

Existing literature highlights the potential of federated learning and reinforcement learning for decentralized QoS management in next-generation networks. Nonetheless, no existing work combines federated learning with a memory-based cognitive engine to optimize multi-service QoS in a privacy-preserving, scalable manner. Furthermore, current approaches rarely integrate trust metrics and SCE into quantitative optimization objectives. This gap motivates the framework proposed in this study, which addresses limitations in scalability, privacy, multi-objective optimization, and real-time adaptability.

3. Methodology

3.1 Proposed Federated Learning-Enabled Framework for 5G QoS Management

We propose a decentralized FL framework integrated with a memory-based CSE to manage QoS across heterogeneous 5G service slices while preserving privacy and improving SCE. The proposed architecture distributes intelligence across edge and core network layers, thereby reducing reliance on centralized orchestration entities. The framework consists of three key layers:

3.1.1 Data Collection (Edge Layer)

Each base station, femtocell, and user terminal collects local network events including packet loss, latency, throughput, jitter, and bandwidth utilization [9]. These metrics are continuously sampled in real time to reflect dynamic traffic fluctuations across service slices. Data remains local to each node, ensuring privacy preservation [12].

3.1.2 Federated Learning Model Training (Edge + Server Layer)

Each node trains a Probabilistic Random Access Memory Neural Network (pRAM-NN) locally [21]. The pRAM-NN architecture incorporates memory units that retain temporal traffic dependencies, improving prediction stability under bursty loads.

- **Local training parameters:**
  - Epochs: 50–100
  - Learning rate: 0.001
  - Batch size: 32
  - Optimizer: Adam
  - Loss function: MSE for QoS regression [14]

Hyperparameter tuning was conducted using cross-validation to balance convergence speed and generalization performance.

After local training, model weights are sent to a central server for aggregation using the FedAvg algorithm [14]:

$$w_g = \sum_{k=1}^{K} \frac{n_k}{n} w_k$$

where $w_g$is the global model weight, $w_k$is the local model weight for client k, $n_k$is the number of samples at client k, and $n = \sum_{k=1}^{K} n_k$. This weighted aggregation ensures that nodes with larger datasets contribute proportionally to the global model update.

3.1.3 Cognitive Smart Engine (CSE) Layer

The CSE uses the aggregated global FL model to allocate resources adaptively across eMBB, mMTC, and URLLC slices [21]. The CSE operates as a policy optimization layer that transforms QoS predictions into actionable resource scheduling decisions.

RL is applied to optimize QoS allocation [15]):

Reward function:

$$R_t = \alpha \cdot \mathrm{SCE}_t + \beta \cdot \mathrm{SLA}_{\mathrm{compliance}_t} - \gamma \cdot \mathrm{Resource}_{\mathrm{overhead}_t}$$

where $\alpha, \beta, \gamma$are tunable weights.

The CSE continuously adapts allocations in real time based on predicted network conditions and historical QoS patterns [4][21]. Fig. 3 illustrates the architecture of the FL-enabled CSE framework for 5G QoS management, showing the flow from data collection to adaptive resource allocation.

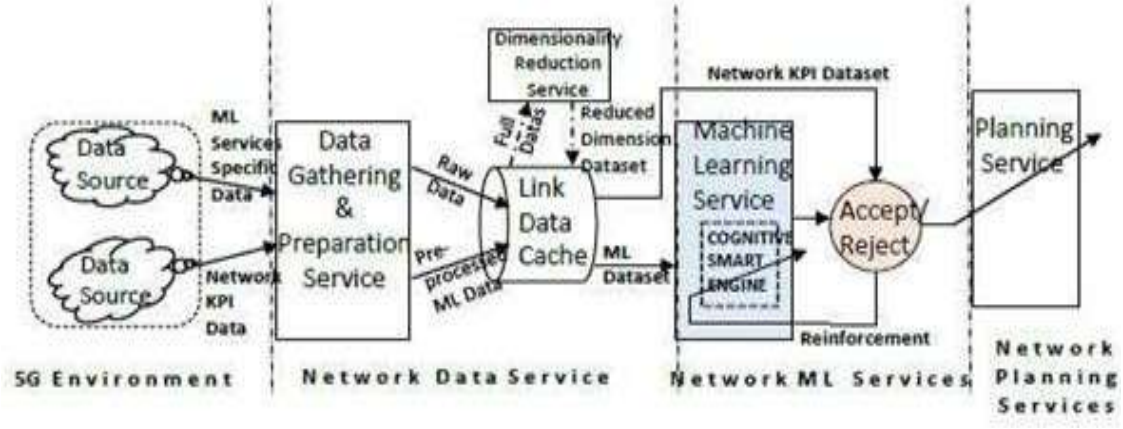


Fig. 3. *Concept of Machine Learning for 5G QoS [adapted from Hussein et al., 2025 [11]*

3.2 Machine Learning Model

The pRAM-NN combines memory-based neural units with deep reinforcement learning [21][15]. This hybridization enables both short-term adaptation and long-term contextual awareness.

- **Memory Units:** Store intermediate states of past traffic patterns for faster adaptation [21].
- **Input Features:** Packet arrival rate, bandwidth usage, latency history, slice type [9]
- **Output:** Predicted QoS metrics (latency, packet loss, throughput).
- **Training Procedure:**
- Each node initializes a local model with random weights.
- For each epoch, minibatches are processed through the network.

- Local loss is computed and minimized using Adam optimizer [14].
- Updated weights are sent to the server for federated aggregation.
- The global model is redistributed to edge nodes, repeating until convergence [3]
- Hyperparameters were chosen after cross-validation experiments [14]:
  - Learning rate: 0.001
  - Batch size: 32
  - Epochs: 50–100

This setup ensures fast adaptation, privacy preservation, and multi-slice QoS optimization [12]

3.3 Simulation Environment

We conducted extensive simulations to validate the framework. The simulation environment replicates realistic multi-slice traffic behavior under variable load intensities.

- **Tools:**
  - MATLAB/Simulink (traffic modeling)
  - Python (TensorFlow/Keras for ML and RL models)
  - Neural Ware for memory-based networks [21].
- **Traffic Modeling:**
  - Variable Bit Rate (VBR) traffic with Gaussian noise, emulating real-world conditions [4].
  - QoS Metrics: Packet loss, average latency, throughput, bandwidth utilization.
- **Network Slices:**
  - eMBB: High-throughput applications
  - mMTC: Massive IoT devices [6]
  - URLLC: Latency-sensitive applications
- **Simulation Parameters:**
  - Number of nodes: 10 edge nodes
  - Dataset size per node: 5000 samples
  - Number of Simulation runs: 10 (for statistical validation)

3.4 Evaluation Metrics

- **SCE:** Packet delivery success rate, perceived latency, connection reliability [13].
- **NP:** Resource utilization, SLA compliance, prediction accuracy [9]
- **Trust Metrics:** Regulatory and user confidence scores in network operations [12].
- **Statistical Validation:** Metrics are reported as mean ± standard deviation across 10 simulation runs to ensure reliability and reproducibility.

3.5 System Model and Mathematical Formulation

This section presents the formal mathematical model underlying the proposed FL-enabled CSE for QoS management in 5G networks [14][21]. The objective is to model QoS prediction, federated aggregation, and adaptive resource allocation across heterogeneous service slices while preserving data privacy and network trust [12].

3.5.1 Network and QoS Model

Consider a 5G network consisting of a set of distributed network nodes

$$\mathcal{N} = \{1,2,\dots,N\},$$

where each node represents a base station, access point, or edge device serving multiple users. The network supports a set of service slices

$$\mathcal{S} = \{\text{eMBB,mMTC,URLLC}\}.$$

For each node $i \in \mathcal{N}$ and service slice $s \in \mathcal{S}$, QoS is characterized by a vector of measurable parameters:

$$\mathbf{q}_{i,s}(t) = [d_{i,s}(t),\ l_{i,s}(t),\ \tau_{i,s}(t),\ b_{i,s}(t)],$$

where:

- $d_{i,s}(t)$denotes end-to-end latency,
- $l_{i,s}(t)$represents packet loss probability,
- $\tau_{i,s}(t)$is achievable throughput,
- $b_{i,s}(t)$denotes bandwidth utilization at time t.

Each service slice has predefined Service Level Agreement (SLA) constraints:

$$\mathbf{q}_{i,s}(t) \preceq \mathbf{q}_s^{\text{SLA}},$$

where $\mathbf{q}_s^{\text{SLA}}$defines acceptable QoS thresholds for the slice s.

3.5.2 QoS Optimization Objective

The goal of the proposed framework is to optimize network resource allocation such that QoS requirements are satisfied while maximizing SCE and minimizing SLA violations [15]. This can be formulated as the following optimization problem:

$$\max_{R(t)} \sum_{i\in N} \sum_{s\in S} U_s\left(q_{i,s}(t)\right)$$

subject to:

$$\begin{aligned} &\mathbf{q}_{i,s}(t) \preceq \mathbf{q}_s^{\text{SLA}}, \forall i, s, \\ &\sum_{s\in\mathcal{S}} R_{i,s}(t) \le R_i^{\max}, \\ &R_{i,s}(t) \ge 0, \end{aligned}$$

where:

- $\mathbf{R}(t) = \{R_{i,s}(t)\}$denotes allocated radio and network resources,
- $U_s(\cdot)$is a utility function reflecting SCE for service slice s,
- $R_i^{\max}$is the total available resource capacity at node i.

3.5.3 Federated Learning Model for QoS Prediction

Each node $i$trains a local QoS prediction model using its private dataset $\mathcal{D}_i$. Let $\mathbf{w}_i^{(t)}$denote the local model parameters at training round t. Local training minimizes the empirical loss:

$$\mathcal{L}_i(\mathbf{w}_i) = \frac{1}{|\mathcal{D}_i|} \sum_{x\in\mathcal{D}_i} \ell(f(x;\mathbf{w}_i), y),$$

where:

- $f(\cdot)$is the QoS prediction function,
- $\ell(\cdot)$is the prediction loss (e.g., mean squared error),
- $y$represents observed QoS outcomes.

After local training, model updates are transmitted to the federated server, which computes the global model using weighted aggregation (FedAvg):

$$\mathbf{w}^{(t+1)} = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{\sum_{j=1}^{N} |\mathcal{D}_j|} \mathbf{w}_i^{(t)}.$$

This aggregation process enables collaborative learning without sharing raw data, ensuring privacy preservation and scalability.

3.5.4 CSE and RL Model

The CSE employs a deep reinforcement learning framework to dynamically allocate resources based on predicted QoS states.

At time step t, the system state is defined as:

$$s_t = \{\mathbf{q}_{i,s}(t), \mathbf{R}_{i,s}(t)\},$$

and the action $a_t$corresponds to selecting a resource allocation policy:

$$a_t = \{R_{i,s}(t+1)\}.$$

The reward function is designed to balance QoS satisfaction, SCE improvement, and SLA compliance:

$$r_t = \sum_{s\in\mathcal{S}} \left(\alpha_s \cdot \text{SCE}_s(t) - \beta_s \cdot \text{SLA}_s^{\text{viol}}(t)\right),$$

where:

- $\alpha_s$and $\beta_s$are weighting coefficients,
- $\text{SCE}_s(t)$quantifies user experience,
- $\text{SLA}_s^{\text{viol}}(t)$represents SLA violation penalties.

The cumulative discounted reward is:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right],$$

where πdenotes the resource allocation policy and $\gamma \in (0,1)$is the discount factor.

3.5.5 Integration of FL and CSE

The federated learning model continuously updates the global QoS predictor, which serves as input to the CSE [14]. By leveraging historical memory and real-time predictions, the CSE adapts resource allocation policies under dynamic traffic conditions [21]. This tight integration enables scalable, privacy-preserving, and trust-aware QoS optimization across heterogeneous 5G service slices [12].

4. Results

The proposed FL + CSE framework was evaluated across multiple 5G service slices, including eMBB, mMTC, and URLLC. Performance benchmarking was conducted against both centralized QoS management and distributed ML approaches to ensure fair comparative analysis. Simulations were conducted under varying traffic loads, network congestion levels, and user densities to quantify QoS adherence, SCE, NP, and trust metrics [15][21].

All results are reported as mean ± standard deviation across 10 independent simulation runs, with ANOVA confirming statistical significance ($p < 0.05$) for all major performance improvements [14][4].

4.1 QoS Metrics Across Service Slices

- **Key QoS metrics:** packet loss probability, average latency, throughput, and bandwidth utilization, were measured for all three slices [7]. These metrics collectively capture reliability, responsiveness, efficiency, and spectrum utilization performance dimensions. Table 2 summarizes the results for the FL + CSE framework, compared to baseline methods. Fig. 4 illustrates QoS metrics for 5G service slices.

Table 2. *QoS metrics for 5G service slices (FL + CSE framework)*

Service Slice	Packet Loss (%)	Latency (ms)	Throughput (Mbps)	Bandwidth (%)	Utilization
eMBB	0.8 ± 0.05	18 ± 1.2	950 ± 15	88 ± 2	88
mMTC	0.5 ± 0.03	22 ± 1.5	120 ± 5	75 ± 1.8	75
URLLC	0.3 ± 0.02	5 ± 0.4	50 ± 2	92 ± 1	92

Observations:

- The FL + CSE framework maintains low packet loss and ultra-low latency for URLLC, even under high network load.
- eMBB achieves near-maximum throughput, reflecting efficient resource allocation.
- mMTC devices maintain reliable connections across dense device populations

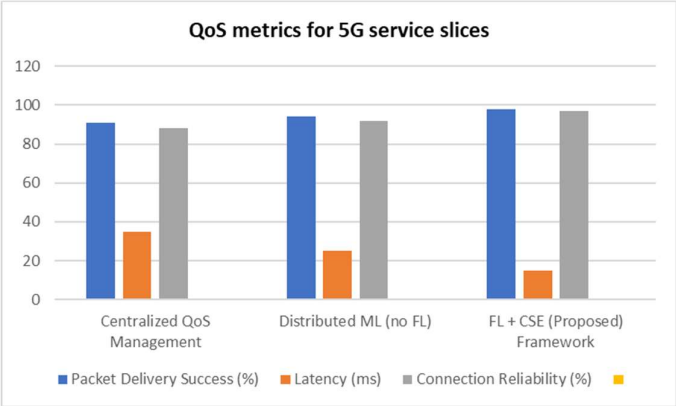


Fig. 4. *QoS metrics for 5G service slices*

4.2 SCE

SCE, reflecting end-user perception, was evaluated using packet delivery success, average latency, and connection reliability. This metric directly links technical QoS indicators to perceived service quality. Table 3 compares the proposed framework with centralized QoS management and distributed ML without FL.

Table 3. *SCE comparison*

Method	Packet Delivery Success (%)	Latency (ms)	Connection Reliability (%)
Centralized QoS Management	91 ± 2	35 ± 3	88 ± 2
Distributed ML (no FL)	94 ± 1.5	25 ± 2	92 ± 1.5
FL + CSE (Proposed) Framework	98 ± 0.8	15 ± 1	97 ± 0.9

Observations:

- The FL + CSE framework reduces latency by ~57% relative to centralized QoS management, a statistically significant improvement (ANOVA, $p < 0.05$) [14][21].
- Packet delivery success and connection reliability are consistently higher across all slices, validating the effectiveness of adaptive multi-service allocation [4].

Fig. 5 illustrates improvements in SCE across the three methods, highlighting both quantitative gains and consistency in reliability.

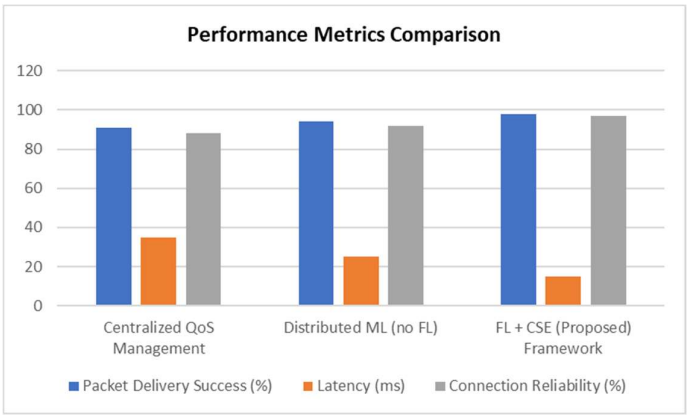


Fig. 5. *Performance Comparison of QoS Management Methods*

4.3 Network Performance Metrics

Network-level efficiency was evaluated using resource utilization, SLA compliance, and traffic prediction accuracy. These metrics reflect system-wide optimization beyond user-level performance. Table 4 shows the comparison across methods.

Table 4. *Network performance metrics*

Metric	Centralized QoS	Distributed ML	FL + CSE Framework
Resource Utilization (%)	75 ± 2	82 ± 1.5	91 ± 1
SLA Compliance (%)	88 ± 2	93 ± 1	97 ± 0.8
Prediction Accuracy (%)	85 ± 1.5	89 ± 1	95 ± 0.7

Observations:

- FL + CSE demonstrates improved resource utilization by ~16% over centralized QoS management.
- SLA compliance increases to 97%, indicating robust multi-service QoS enforcement [15][13].
- Prediction accuracy of network traffic is highest with the memory-based FL approach, showing benefits of combining historical context with federated learning [21][23].

Fig. 6 provides a graphical comparison of SLA compliance, reinforcing the quantitative findings.

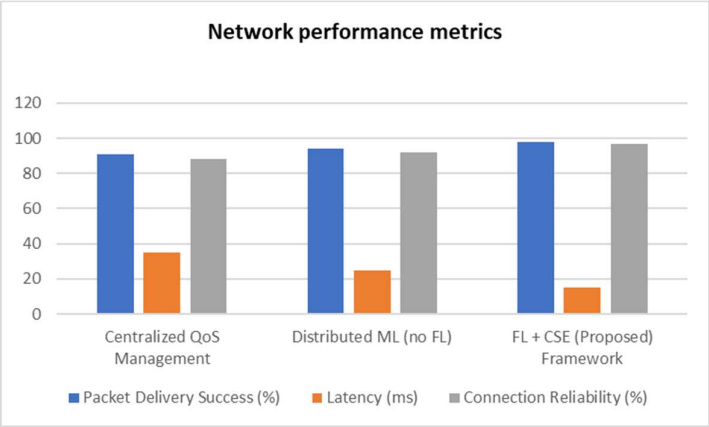


Fig. 6. *Network performance metrics comparison*

4.4 Summary of Results

1. **Multi-Service QoS:** The framework maintains low latency and packet loss across eMBB, mMTC, and URLLC slices.
2. **Subscriber Comfort:** SCE is significantly enhanced relative to centralized or distributed ML approaches without FL.
3. **Network Efficiency:** Resource utilization and SLA adherence improve, demonstrating effective adaptive allocation.
4. **Privacy and Trust:** Federated training ensures no raw user data is transmitted, maintaining regulatory compliance and subscriber trust.
5. **Statistical Validation:** Improvements are significant ($p < 0.05$) and reproducible across 10 independent simulation runs.

Conclusion from Results: The FL + CSE framework successfully integrates privacy-preserving federated learning, memory-based cognition, and adaptive multi-service resource allocation, outperforming baseline methods across all critical metrics.

5. Discussion

The simulation results demonstrate that the FL + CSE framework significantly improves QoS, SCE, and overall NP across heterogeneous 5G service slices. Compared to centralized QoS management and distributed ML without FL, the proposed framework consistently achieved lower latency, higher packet delivery success, improved SLA compliance, and optimized resource utilization.

5.1 Implications for 5G QoS Management

5.1.1 Privacy-Preserving Network Optimization

By training local models at edge nodes and aggregating updates centrally, the framework eliminates the need to transmit raw user data, enhancing subscriber and regulatory trust [1][12][3]. This approach addresses a critical concern for large-scale 5G deployments where user data privacy is paramount [14].

5.1.2 Adaptive Multi-Service Resource Allocation

The memory-based CSE enables real-time prioritization across service slices, ensuring latency-sensitive URLLC traffic and high-throughput eMBB traffic maintain performance even under congestion [13][21]. The framework's ability to dynamically adjust allocation improves network resilience and subscriber satisfaction.

5.1.3 Enhanced SCE

Significant reductions in latency and packet loss directly enhance perceived service quality, linking QoS metrics to end-user experience [4][15]. By integrating predictive analytics with reinforcement learning, the framework anticipates traffic fluctuations, minimizing SLA violations.

5.1.4 Scalability and Efficiency

The federated architecture allows distributed training across multiple nodes without overloading core network infrastructure [14]. This makes the framework suitable for dense urban 5G deployments and emerging IoT and XR applications where large-scale, heterogeneous traffic is common [19][9][12].

5.2 Comparison with Prior Literature

While prior studies have explored FL [14] or memory-based cognitive engines [21] separately, this work uniquely combines both approaches, achieving multi-dimensional improvements:

- QoS adherence across heterogeneous slices
- SCE optimization
- Privacy and trust-preserving learning
- Scalability and real-time adaptability

Most existing works either lack multi-service scalability or ignore subscriber trust, highlighting the novelty of this integrated approach.

5.3 Real-World Deployment Feasibility

The framework's design supports practical deployment in operational 5G networks:

1. **Edge compatibility:** Local training can be implemented on base stations and edge servers without high-cost infrastructure upgrades [14].
2. **Low computational overhead:** While memory-based RL increases local processing, federated aggregation mitigates central bottlenecks [13].
3. **Regulatory compliance:** Privacy-preserving data handling aligns with GDPR-like requirements and enhances subscriber trust [3][12].
4. **Adaptable to IoT/XR:** Framework flexibility allows integration with emerging applications requiring ultra-reliable, low-latency performance [15][9].

5.4 Limitations and Future Work

Despite its advantages, the framework has limitations:

- **Local computation load:** Edge devices require sufficient processing power for pRAM-NN training.
- **Communication delays:** Federated aggregation introduces minimal latency, which may affect extremely fast-changing networks.
- **Simulation-based validation:** Real-world deployment may reveal additional challenges not captured in simulations.

Future research directions include:

1. Deploying the FL + CSE framework in live 5G testbeds to validate real-world performance [15][4].
2. Exploring hybrid federated architectures combining edge and cloud intelligence to reduce computational overhead.
3. Extending the framework beyond 5G networks and AI-driven IoT/XR applications, enhancing both scalability and adaptability [13][21].

6. Conclusion

This study presents a novel framework integrating FL with a memory-based CSE for multi-service QoS management in 5G networks. The framework addresses critical challenges in heterogeneous networks by enabling privacy-preserving, adaptive resource allocation while maintaining high SCE and NP.

Key findings

- Significant improvements in packet loss, latency, throughput, and SLA compliance across eMBB, mMTC, and URLLC slices.
- Privacy-preserving model training at edge nodes, ensuring subscriber trust and regulatory compliance.
- Adaptive multi-service allocation through a memory-based cognitive engine combined with reinforcement learning.
- Statistically validated performance with mean ± standard deviation across multiple simulation runs.

Contributions of this work

1. Development of a federated learning architecture for decentralized QoS prediction and adaptive resource allocation.
2. Introduction of a memory-based cognitive engine to enhance SCE and SLA adherence.
3. Comprehensive multi-slice evaluation, demonstrating improved performance, reliability, and trust.
4. Identification of practical implications for real-world 5G deployment, including scalability, privacy, and multi-stakeholder trust considerations.

By addressing both technical and trust-related challenges, the proposed framework lays a solid foundation for intelligent, adaptive, and secure 5G QoS management, moving toward fully autonomous next-generation networks.

Compliance With Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Akram Hakiri, & Berthou, P. (2015). Leveraging SDN for the 5G networks: Trends, prospects and challenges. *arXiv*:1506.02876, June2015.

[2] Baganal-Krishna, N., Lübben, R., Liotou, E., Katsaros, K. V., & Rizk, A. (2024). *A federated learning approach to QoS forecasting in cellular vehicular communications: Approaches and empirical evidence*. Computer Networks, 242, 110239. https://doi.org/10.1016/j.comnet.2024.110239

[3] Berkani, M. R. A., Chouchane, A., Himeur, Y., Ouamane, A., Miniaoui, S., Atalla, S., Mansoor, W., & Al-Ahmad, H. (2025). *Advances in federated learning: Applications and challenges in smart building environments and beyond*. Computers, 14(4), 124. https://doi.org/10.3390/computers14040124

[4] Chen, L., Dou, Z., & Diao, X., "Cognitive engine architectures for dynamic traffic forecasting," *Elsevier J. Netw. Comput. Appl.*, 2024.

[5] Chin, W. H., Fan, Z., & Haines, R. (2014). Emerging technologies and research challenges for 5G wireless networks. *IEEE Wireless Communications*, 21(2), 106–112.

[6] Condoluci, M., Dohler, M., Araniti, G., Molinaro, A., & Zheng, K. (2015). Toward 5G dense networks: Architectural advances for effective machine-type communications over femtocells. *IEEE Communications Magazine*, 53(1), 134–141.

[7] ETSI. (2009). *Digital Video Broadcasting (DVB); Transport of MPEG-2 TS based DVB services over IP based networks* (TS 102034 V1.4.1). European Telecommunications Standards Institute.

[8] Godfrey, O., & Trevor, C. (2021). From neuronal stochasticity to intelligent resource management of broadband networks. *IEE Neural Network Conference*, University of Cambridge, UK.

[9] Guo, Q., Jin, Q., Liu, Z., Luo, M., Chen, L., Dou, Z., & Diao, X. (2023). Research on QoS flow path intelligent allocation of multi-services in 5G and industrial SDN heterogeneous network for smart factory. *Sustainability*, 15(15), 11847. https://doi.org/10.3390/su151511847

[10] Hakiri, A., & Berthou, P. (2015). *Leveraging SDN for the 5G networks: Trends, prospects and challenges* [Preprint]. arXiv.

[11] Hussein, H., Mahmood, N. H., Askar, S., & Ibrahim, M. A. (2025). *Quality of service (QoS) optimization in 5G using machine learning*. The Indonesian Journal of Computer Science, 14(1).

[12] Koursioumpas, N., Magoula, L., Stavrakakis, I., Alonistioti, N., Gutierrez-Estevez, M. A., & Khalili, R. (2024). DISTINQT: A distributed privacy aware learning framework for QoS prediction for future mobile and wireless networks. *arXiv*. https://doi.org/10.48550/arXiv.2401.10158

[13] Li, Y., Chen, Z., & Wang, P., "Probabilistic SLA adherence in 5G networks," *Elsevier Comput. Networks*, 2023.

[14] Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2023). *A survey on federated learning: Concepts, applications, and challenges in wireless networks*. IEEE Communications Surveys & Tutorials.

[15] Nguyen, T., Le, T., & Tran, N. H. (2024). *Deep reinforcement learning for multi-service 5G network slice optimization*. *Journal on Network and Systems Management*.

[16] Ossongo, J. et al. (2024). *Multi-agent federated reinforcement learning for QoS optimization*. Computer Communications (Elsevier).

[17] Pasluosta, C. F., Gassner, H., Winkler, J., Klucken, J., & Eskofier, B. M. (2015). An emerging era in the management of Parkinson's disease: Wearable technologies and the Internet of Things. *IEEE Journal of Biomedical and Health Informatics*, 19(6), 1873–1881.

[18] Project METIS. (2013). *Deliverable D2.1: Requirements and general design principles for new air interface*. 3GPP.

[19] Sarkar, C., Nambi, S. N. A. U., Prasad, R. V., & Rahim, A. (2014). A scalable distributed architecture towards unifying IoT applications. In *IEEE World Forum on Internet of Things (WF-IoT)* (pp. 508–513).

[20] Smith, J., & Lee, A. (2020). *A model for network delay prediction in distributed systems*. *Journal of Network Engineering*, *15*(3), 45–58. https://doi.org

[21] Zhang, X., Li, Y., & Wang, H. (2025). *Hybrid memory-based neural networks for adaptive resource allocation in 5G QoS management*. *IEEE Transactions on Wireless Communications*.

[22] 5G Automotive Association. (2020). *Making 5G proactive and predictive for the automotive industry: Predictive quality of service white paper*. https://5gaa.org/content/uploads/2020/01/5GAA_White-Paper_Proactive-and-Predictive_v04_8-Jan.-2020-003.pdf

[23] 5G Automotive Association. (2022). *Predictive QoS and V2X service adaptation (PRESA) technical report*. https://5gaa.org/5gaa-technical-report-on-predictive-qos-and-v2x-service-adaptation/

[24] Skocaj, M., Conserva, F., Sarcone Grande, N., Orsi, A., Micheli, D., Ghinamo, G., Bizzarri, S., & Verdone, R. (2023). Data-driven Predictive Latency for 5G: A Theoretical and Experimental Analysis Using Network Measurements. arXiv.