Optimizing Edge Gaming Slices through an Enhanced User Plane Function and Analytics in Beyond-5G Networks

Bruno Marques da Silva[1], **Larissa Ferreira Rodrigues Moreira**[1], **Flávio de Oliveira Silva**[2] **and Rodrigo Moreira**[1]

[1]Institute of Exact and Technological Sciences – Federal University of Viçosa (UFV)
Rio Paranaíba – MG – Brazil

[2]Department of Informatics – School of Engineering
University of Minho (UMinho) – Braga – Portugal

{bruno.silva63, rodrigo, larissa.f.rodrigues}@ufv.br, flavio@di.uminho.pt

Abstract. *The latest generation of games and pervasive communication technologies poses challenges in service management and Service-Level Agreement compliance for mobile users. State-of-the-art edge-gaming techniques enhance throughput, reduce latency, and leverage cloud computing. However, further development of core functions such as the User Plane Function (UPF) is needed for non-intrusive user latency measurement. This paper proposes a closed-loop architecture integrating the Network Data Analytics Function (NWDAF) and UPF to estimate user latency and enhance the 5G control plane by making it latency-aware. The results show that embedding an artificial intelligence model within NWDAF enables game classification and opens new avenues for mobile edge gaming research.*

1. Introduction

Advancements in Fifth-generation of Mobile Telecommunications Technology (5G) networks and the unprecedented capacity of Graphics Processing Unit (GPU) have opened up support for bold network metrics to meet the demands of business verticals such as Virtual Reality (VR), Augmented Reality (AR), and entertainment applications such as online gaming [Shankar 2024]. Cloud gaming, where a server streams games to users or devices, presents challenges due to network dynamism, requiring high throughput for streaming Key Performance Indicators (KPIs) and low latency for satisfactory Quality of experience (QoE) [Soares et al. 2024].

Artificial Intelligence (AI) plays a crucial role in network slicing management and orchestration, enabling precise service customization for resource-intensive applications [Moreira et al. 2023, Rodrigues Moreira et al. 2024]. Enhancing Edge gaming in 5G remains challenging due to the need for coordinated interventions across Radio Access Network (RAN), User Plane Function (UPF), and cloud services [Soares et al. 2024]. In this context, AI supports Quality of Service (QoS) assurance and Service-Level Agreement (SLA) compliance in dynamic environments [Kougioumtzidis et al. 2024], allowing mobile network control plane mechanisms to enforce strict performance metrics for deployed network slices, particularly in online gaming.

State-of-the-art edge gaming approaches, including Multi-access Edge Computing (MEC)-based methods, wireless network enhancements, and traffic engineering for

QoS support [Soares et al. 2024, Shankar 2024, Kougioumtzidis et al. 2024], do not incorporate the Network Data Analytics Function (NWDAF) function from 3rd Generation Partnership Project (3GPP) Release 16, which provides analytics in the 5G core network. This paper introduces UPF instrumentation with a user-space filter to measure User Equipment (UE) latency and report it to NWDAF for core network analysis and potential SLA improvements. The main contributions are i) UPF instrumentation with user-space filters for slice latency estimation based on Tunnel Endpoint Identifier (TEID), ii) empirical evaluation using a real dataset, and iii) performance assessment of AI techniques in this domain.

The remainder of this paper is organized as follows: Section 2 reviews prior work on cloud gaming challenges; Section 3 outlines the proposed method; Section 4 describes the testbed and technologies used; Section 5 analyzes the results, insights, and lessons learned; and Section 6 presents conclusions and future work.

2. Related Work

[Slivar et al. 2019] addressed the optimization of resource allocation for multiple cloud gaming users sharing a bottleneck link in 5G networks by employing QoE-aware algorithms based on subjective QoE models; their regression analysis utilized Mean Opinion Score (MOS) scores from games such as Serious Sam 3, Hearthstone, and Orcs Must Die! Unchained, derived from controlled laboratory experiments.

[Zhang et al. 2019] developed the EdgeGame, a framework leveraging mobile edge computing to address high latency and bandwidth consumption in cloud gaming using a deep reinforcement-learning-based algorithm for adaptive bitrate control. The system optimizes QoE under dynamic network conditions.

[Baena et al. 2023] proposed a comprehensive dataset containing quality indicators to evaluate video streaming and cloud gaming services' End-to-End (E2E) performance over 5G networks, utilizing a regression approach to estimate E2E service metrics based on network parameters.

[Rossi et al. 2024] evaluated three objective QoE prediction models for mobile cloud gaming, leveraging linear, polynomial, and nonlinear regression to address the impact of QoS factors including Round-trip time (RTT) and the models were trained and validated on a publicly available dataset derived from controlled subjective tests.

[Carvalho et al. 2024] employed transfer learning to address the challenge of cross-domain QoE estimation in cloud gaming services, focusing on adaptation from wired to mobile 5G networks. Their regression-based model significantly reduced the Mean Squared Error (MSE) by leveraging a dataset of subjective QoE assessments collected under varying network conditions.

[Soares et al. 2024] proposed an expanded stacking learning model that integrates datasets from wired and mobile network contexts, focusing on wireless networks (5G). The study employed a regression approach and utilized a merged dataset with 3,323 instances from 88 players, combining features from different gaming environments to effectively predict QoE.

In contrast to the previously mentioned works, our proposal advances the state-of-the-art by introducing a latency-aware closed-loop architecture within the 5G control

plane. By integrating NWDAF and UPF, we enable nonintrusive latency measurement and leverage AI for real-time game classification and latency forecasting. This innovative approach ensures SLA compliance, enhances edge-gaming service management, and provides a robust solution to address the complexities of modern gaming environments.

3. Proposed Method

This paper proposes a non-intrusive method for user service evaluation, estimating gaming user experience quality solely through network infrastructure analysis. The method calculates the latency within a specific packet flow window passing through the N3 interface of UPF, as illustrated in Figure 1.

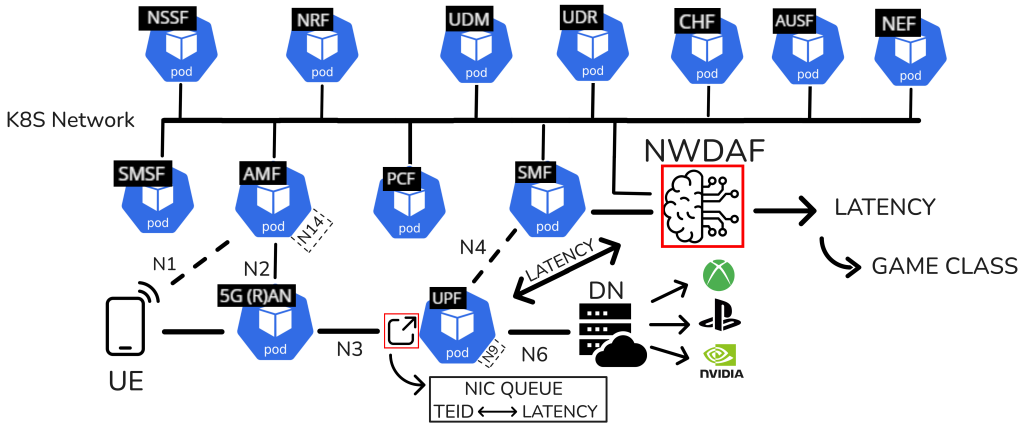


Figure 1. Proposed B5G Architecture for Enhanced Edge Gaming.

The functions of the 5G core are represented at the top, including Network Slice Selection Function (NSSF), Network Repository Function (NRF), Unified Data Management (UDM), Unified Data Repository (UDR), Core Charging Function (CHF), Authentication Server Function (AUSF), Short Message Service Function (SMSF), Access and Mobility Management Function (AMF), Policy Control Function (PCF), Session Management Function (SMF), and Network Exposure Function (NEF), which provide support for session control, policy management, traffic forwarding, and user authentication.

Our method involves instrumenting UPF with a user-space latency monitor that measures the temporal offset of packets for each TEID in each slice. The latency of each received packet was recorded and submitted to NWDAF for analysis. The architecture envisions the NWDAF notifying the SMF about the quality of service perceived by the user, based on the game class consumed by the UE and whether intervention is needed. The framework integrates pretrained Machine Learning (ML) models, including K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Long Short-Term Memory (LSTM), and CatBoost.

Figure 2 illustrates the latency monitor operating on the N3 interface of UPF. The approach analyzes the UPF pod interface, capturing packets using specific filters to extract the TEID, a 32-bit field in the General Packet Radio Service Tunnelling Protocol (GTP) header uniquely identifying the tunnel endpoint. Latency is estimated from the arrival and return timestamps of the packets. Based on this data, NWDAF can classify the game by identifying latency patterns.

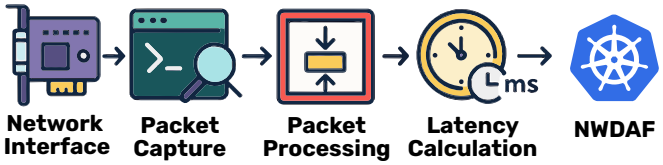


Figure 2. Time-Shift Latency Estimation for Edge Gaming.

Time-shift latency estimation measurement involves observing packet timestamps at an UPF node without introducing additional traffic. If a packet arrives at the UPF at t_{in} and departs at t_{out}, the latency is expressed as:

$$L = t_{\text{out}} - t_{\text{in}}$$

For bidirectional communication, the total latency can be calculated as:

$$L_{\text{total}} = (t_{\text{out}}^{\text{request}} - t_{\text{in}}^{\text{request}}) + (t_{\text{out}}^{\text{response}} - t_{\text{in}}^{\text{response}})$$

As shown in Figure. 2, we construct a filter in the user-space of the UPF to associate the estimated latency with the TEID, enabling us to estimate the latency for different User Equipments (UEs) that may be registered in the 5G core.

Furthermore, our proposed framework assumes an NWDAF containing an Application Programming Interface (API) of pre-trained ML models to instruct the SMF regarding the state of the UEs sessions. Based on the estimated latencies, NWDAF identified the type of game played by UE. If service degradation is detected, the SMF can take corrective actions, such as reconfiguring resource allocation, suggesting routing policy changes, or dynamically adapting network parameters to optimize the user experience.

4. Experimental Setup

This paper instantiates a Fabric testbed virtual machine with 32 GB RAM and 16 vCPUs, running a Kubernetes 1.28 cluster. The free5GC services, including control plane and user plane components, are deployed on this cluster. For the proof of concept, we used a dataset comprising 69,395 instances and 16 features, representing characteristics extracted from various games [Hassanein et al. 2025]. As a classification problem, the target variable corresponds to the game category, consisting of three classes: League of Legends (LOL), Teamfight Tactics (TFT), and Valorant (VAL).

Features encompass numeric and categorical attributes, including source, destination, latitude, and longitude, which capture key aspects of game behavior and performance. These features were structured during preprocessing to enable statistical analysis and model processing for integration into the framework.

5. Results and Discussion

We evaluated the UE-experienced latency by analyzing the time-shifted packet flow through the N3 interface of the UPF. Baseline latency, defined as the UE-perceived RTT when transmitting data, was measured while inducing a sinusoidal synthetic load (1 ms

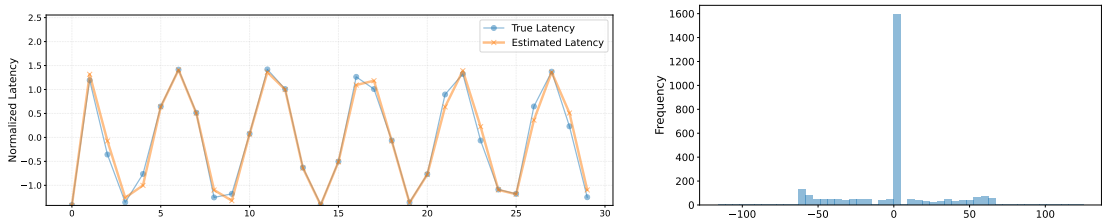
to 600 ms over a 30-second cycle) on the packet recipient. Simultaneously, our filter captured timestamps of packets transiting the N3 interface with a specific TEID to estimate intermediate latency.

Table 1 shows our non-intrusive method effectively estimates UE latencies, with low normalized errors (MSE: 0.019, Mean Absolute Error (MAE): 0.085) and a high R^2 (0.980). Despite a higher normalized Mean Absolute Percentage Error (MAPE) of 25.090 in some cases, the original MAPE of 6.352 highlights the model's practical adequacy, with an average error of ≈ 6.352 between actual and estimated latencies via the N3 interface.

Table 1. UPF Latency Estimation Performance.

Metric	Value
MSE (Normalized)	0.019
MAE (Normalized)	0.085
MAPE (Normalized)	25.090
R2 Score (Normalized)	0.980
MAPE (Original)	6.352

Figure 3a depicts a 10-minute time series sample comparing the latency experienced by the UE with the estimated latency in the cluster where the UPF was deployed. This reinforces the approach's promise and accuracy, as Figure 3b highlights a significant accumulation of estimation errors near zero.



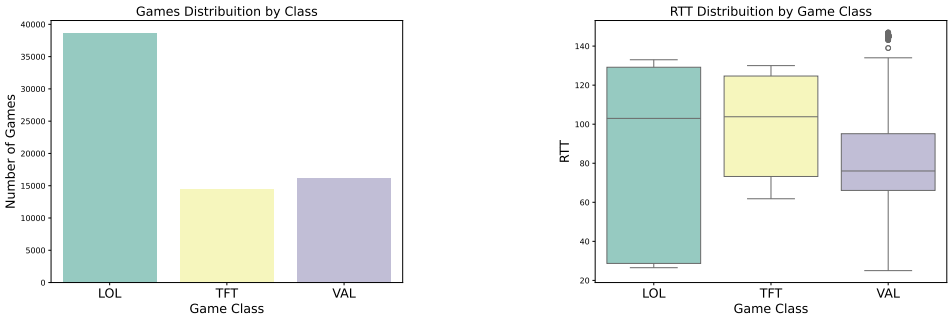
(a) Sample of Real and Estimated Latency. (b) Histogram of Errors.

Figure 3. Estimation Analyze.

We analyzed the dataset's quality to train the ML model. As shown in Figure 4a, the class distribution is unbalanced: LOL has the largest share with nearly 40,000 occurrences, followed by VAL with approximately 17,000, and TFT with approximately 15,000 games. Figure 4b illustrates the RTT distribution for LOL, TFT, and VAL, enabling the analysis of latency variability and its potential impact on player experience.

Analyzing the interquartile range (IQR), which represents central data dispersion, LOL exhibits the highest variability, indicating significant oscillations in network response times and potential challenges for our proof-of-concept, whereas TFT shows a smaller IQR, suggesting more stable latency. While VAL displays the most significant variability across the whole data range, LOL has the highest RTT variability in the central distribution, potentially impacting gameplay predictability. Table 2 summarizes the average performance metrics of the five classification algorithms over 10 runs, including Accuracy, Precision, Recall, and F1-Score.

The CatBoost algorithm achieved the best overall performance, with an accuracy of 0.9483 and an F1-Score of 0.9477, surpassing all other models across metrics. At the same time, Random Forest performed similarly (accuracy: 0.9481, F1-Score: 0.9479) but



(a) How the Games are Distributed by Class. (b) RTT Distribution.

Figure 4. Comparison of game distribution and RTT distribution.

Table 2. Average Performance Metrics.

Algorithm	Accuracy	Precision	Recall	F1-Score
KNN	0.9446	0.9441	0.9446	0.9443
RF	0.9481	0.9478	0.0981	0.9479
DT	0.9430	0.9429	0.9430	0.9430
CatBoost	0.9483	0.9476	0.9483	0.9477
LSTM	0.9077	0.9084	0.9077	0.9057

exhibited a low Recall (0.0981), indicating possible class imbalance. Figure 5 shows that decision tree-based models (CatBoost, DT, RF) and KNN demonstrated high stability and consistent accuracy with low variability, whereas LSTM exhibited significant variability, marked by a wide confidence interval and multiple outliers, emphasizing its sensitivity to hyperparameters and data structure; thus, decision tree-based models are more reliable for predictable applications, while LSTM may require adjustments to improve stability.

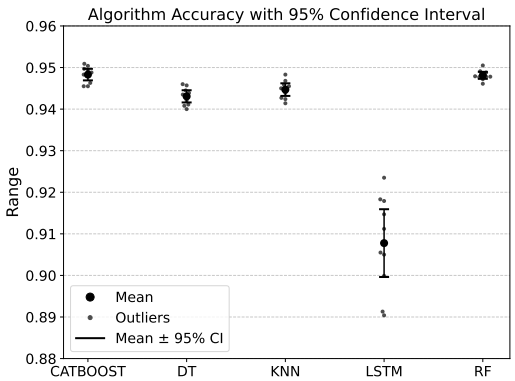
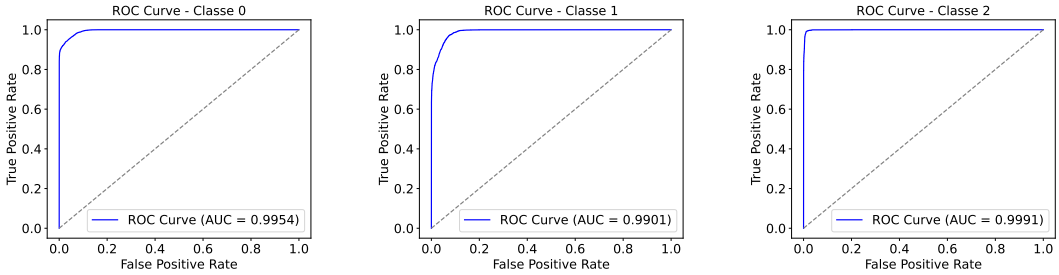


Figure 5. Accuracy Confidence Interval.

Figure 6 shows the Receiver Operating Characteristic (ROC) curves for the three classes analyzed by the CatBoost model: (a) LOL, (b) TFT, and (c) VAL. These curves evaluate the binary and multiclass classification models and represent the true positive rate (sensitivity) against the false positive rate.

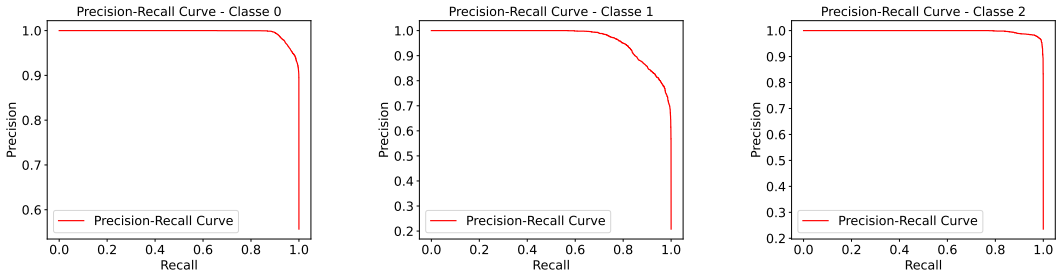
The curves in Figure 7 exhibit strong class distinction near the upper-left corner, indicating high performance, with Area Under the Curve (AUC) values exceeding 0.99 for all classes, demonstrating the model's accuracy and robustness. An AUC near 1.0 signifies excellent discriminatory capability, while values around 0.5 indicate chance-

level performance. The precision-recall curves for the three analyzed classes—(a) LOL, (b) TFT, and (c) VAL)—further assess the CatBoost model's performance in imbalanced scenarios, providing insight into the trade-off between precision (correct positive predictions) and recall (accurate identification of positive instances).



(a) LOL (b) TFT (c) VAL

Figure 6. ROC Curve of CatBoost



(a) LOL (b) TFT (c) VAL

Figure 7. Precision-Recall Curve of CatBoost

All curves exhibited high precision across varying recall levels, reflecting the strong classification accuracy of the model. Minor drops in precision occur only at extreme recall values, which is typical for highly reliable classification. These findings highlight the suitability of our framework for proposing user-plane interventions to enhance the UE quality of experience in online gaming sessions.

6. Concluding Remarks

This paper presents a method to enhance edge gaming in 5G networks by instrumenting the UPF with a user-space filter capable of estimating UE latency using a time-shift approach. While existing approaches focus on improving cloud gaming across various network segments, our method introduces an innovative closed-loop framework between the UPF and NWDAF, enabling real-time latency estimation and creating new opportunities for enforcing QoS policies.

Future work will enhance the SMF and PCF to dynamically adapt session parameters based on user experience, enabling real-time feedback integration and optimization for latency-sensitive applications. We also aim to evaluate performance in wired and hybrid setups, and explore AI-driven methods for predictive network adjustments and adaptive resource allocation in mobile edge environments.

Acknowledgments

We acknowledge the financial support of the FAPESP MCTIC/CGI Research project 2018/23097-3 and FAPEMIG (Grant APQ00923-24). We also thank the FCT – Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/Centro ALGORITMI (ALGORITMI/UM) for partially supporting this work.

References

Baena, C., Peñaherrera-Pulla, O. S., Camacho, L., Barco, R., and Fortes, S. (2023). Video Streaming and Cloud Gaming Services Over 4G and 5G: A Complete Network and Service Metrics Dataset. *IEEE Communications Magazine*, 61(9):154–160.

Carvalho, M., Soares, D., and Fernandes Macedo, D. (2024). QoE Estimation Across Different Cloud Gaming Services Using Transfer Learning. *IEEE Transactions on Network and Service Management*, 21(6):5935–5946.

Hassanein, A., Hashemi, M. R., and Shirmohammadi, S. (2025). Gaming And Video Streaming Traffic for 5G Research.

Kougioumtzidis, G., Vlahov, A., Poulkov, V. K., Lazaridis, P. I., and Zaharis, Z. D. (2024). QoE Prediction for Gaming Video Streaming in O-RAN Using Convolutional Neural Networks. *IEEE Open Journal of the Communications Society*, 5:1167–1181.

Moreira, R., Martins, J. S. B., Carvalho, T. C. M. B., and Silva, F. d. O. (2023). On Enhancing Network Slicing Life-Cycle Through an AI-Native Orchestration Architecture. In Barolli, L., editor, *Advanced Information Networking and Applications*, pages 124–136, Cham. Springer International Publishing.

Rodrigues Moreira, L. F., Moreira, R., Martins, E. T., Jansen, V. F., Lima, Y. S., Rodrigues, L. G. F., Travençolo, B. A. N., and Backes, A. R. (2024). Maximizing the Power of Cognitive Services with an AI-as-a-Service Architecture for Seamless Delivery. In *2024 IEEE 13th International Conference on Cloud Networking (CloudNet)*, pages 1–8.

Rossi, H. S., Mitra, K., Åhlund, C., Cotanis, I., Örgen, N., and Johansson, P. (2024). Objective QoE Models for Cloud-Based First Person Shooter Game over Mobile Networks. In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, pages 550–553.

Shankar, V. (2024). Edge AI: A Comprehensive Survey of Technologies, Applications, and Challenges. In *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–6.

Slivar, I., Skorin-Kapov, L., and Suznjevic, M. (2019). QoE-Aware Resource Allocation for Multiple Cloud Gaming Users Sharing a Bottleneck Link. In *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pages 118–123.

Soares, D., Carvalho, M., and Macedo, D. F. (2024). Enhancing Cloud Gaming QoE Estimation by Stacking Learning. *Journal of Network and Systems Management*, 32(3):58.

Zhang, X., Chen, H., Zhao, Y., Ma, Z., Xu, Y., Huang, H., Yin, H., and Wu, D. O. (2019). Improving Cloud Gaming Experience through Mobile Edge Computing. *IEEE Wireless Communications*, 26(4):178–183.