QoE for Interactive Services in 5G Networks: Data-driven Analysis and ML-based Prediction

Stefania Zinno†, Giuseppe Caso‡, Nicola Pasquino†, Alessio Botta†, Anna Brunstrom‡ and Giorgio Ventre†
†DISS, DIETI
Università degli Studi di Napoli Federico II, 80125 Napoli, Italy
Email: alessio.botta@unina.it; stefania.zinno@unina.it; nicola.pasquino@unina.it; giorgio.ventre@unina.it
‡Department of Mathematics and Computer Science
Karlstad University, 65188 Karlstad, Sweden
Email: giuseppe.caso@kau.se, anna.brunstrom@kau.se

***Abstract*—Nowadays, the focus in 5G networks has shifted from Quality of Service (QoS) to Quality of Experience (QoE) characterisation and prediction. As a matter of fact, mobile operators are increasingly interested in measuring and/or predicting QoE Key Performance Indicators (KPIs) on their 5G networks. In this context, a recent methodology by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) allows to characterize the level of *interactivity* achievable by real-time services on 5G networks, by computing a synthetic QoE KPI referred to as *interactivity score* (*i-score*). The *i-score*, defined as the measurable latency, continuity, and reliability of a given service, is computed by using a model that takes into account three QoS KPIs, i.e., packet trip time, jitter, and loss rate. In this paper, aiming at assessing the effectiveness of the ITU-T methodology in characterizing 5G network performance, we analyze a large-scale measurement campaign executed over two commercial 5G Non-Standalone (NSA) deployments in the city of Rome, Italy. During this campaign, traces related to radio coverage and service performance (i.e., the *i-score* and corresponding KPIs needed to compute it) were collected in parallel. Therefore, we use the dataset to characterize the observed *i-score* performance, and demonstrate that it is possible to successfully predict this KPI with machine learning techniques, using radio layer parameters and power measurements. Mobile operators could take advantage of our findings, minimizing the need for time/resource-consuming QoE tests. Ensemble methods in fact achieve an accuracy spanning from 0.79 to 0.83, with Random Forest as one of the best algorithm to predict the *i-score* from radio layer parameters.**

***Index Terms*—5G, Interactivity Score, Machine Learning, Quality of Experience, Quality of Service**

I. Introduction

Nowadays, 5G networks are able to improve their 4G counterpart, offering an average data rate in downlink of $100\,\mathrm{Mbps}$ with peaks up to $20\,\mathrm{Gbps}$, a reliability of 99,9999%, and a full support for high traffic densities of devices [1]. By the end of 2029, 5G is anticipated to become the leading mobile access technology, with 5G mobile subscriptions projected to reach nearly 5.6 billion by that time [2]. The 5G systems standardized by the 3rd Generation Partnership Project (3GPP) are designed to address the Quality of Service (QoS) and Quality of Experience (QoE) requirements of massive Machine Type Communication (mMTC), enhanced Mobile Broadband (eMBB), and Ultra-Reliable Low Latency Communication (URLLC) applications [3]. Evaluating QoE is vitally important, given the enormous reach and widespread acceptance of 5G networks and the promises standards are making.

QoE of a telecommunications service is defined in multiple documents as: *"The degree of satisfaction of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility or enjoyment of the application or service in light of the user's personality and current state"* [4], *"The degree of delight or annoyance of the user of an application or service"* [5], and *"The overall acceptability of an application or service as perceived subjectively by the end- user"* [6].

Due to the nature of QoE and since it strongly depends on network conditions, most approaches attempt to construct QoE models based on network-specific Key Performance Indicators (KPIs). Moreover, Machine Learning (ML) is pictured today as a key tool for predicting QoE [7]. As a matter of fact, ML is being considered for optimizing several 5G and Beyond-5G (B5G) network aspects and operations, particularly when complex problems cannot be addressed using traditional approaches or models [8]. ML is also revolutionizing the development of QoE prediction models, emphasizing the individualized experience of end-users. A survey examining ML-based QoE predictive models for multimedia service quality in extended reality and video gaming applications can be found in [9].

Some network applications, in fact, significantly suffer from radio layer rapidly changing conditions. As an example, online gaming services can be strongly impacted by 5G channel variations, ultimately affecting players' QoE significantly, especially when dealing with cloud gaming. In this case, QoE models for cloud gaming could be constructed to understand how 5G network conditions affect QoE cloud gaming, through practical tests on players' subjective and objective data [10].

In our previous research [11], we measured QoE and QoS relying on Round Trip Time (RTT), which can be seen as a simple way to evaluate how 5G affects users' QoE [12]. We thus developed our custom methodology based on real-time data collection through a smartphone, and adopted ML to predict RTT performance based on 5G radio layer parameters.

Currently, our research has shifted towards a more synthetic parameter, referred to as *interactivity score (i-score)*, defined in a testing methodology recently standardized by the International

978-3-903176-66-9 ©2024 IFIP

Telecommunication Union Telecommunication Standardization Sector (ITU-T) in Recommendation G.1051 [13]. In particular, using measurements collected during so-called *interactivity tests* [14], we aim at studying how channel conditions encountered on 5G networks affect the QoE of real-time services, e.g., cloud gaming applications. In this work, we thus present an in-depth characterization and analysis of the *i-score* [15], and demonstrate that *i-score* prediction can be carried out through ML techniques, which in turn can be easily implemented as part of the Radio Access Network (RAN), e.g., in a 5G gNode-B.

The rest of the paper is structured as follows: we describe the *i-score* model in Section II and present the experimental campaign in Section III. Section IV provides a brief analysis and characterization of the *i-score* perfomance observed during a specific measurement campaign, while Section V introduces the use of ML for *i-score* prediction, and shows the obtained results. Section VI concludes the paper.

II. QoE Testing For Interactive Services: The *i-score* Concept

The definition of performance testing methodologies for interactive services is key towards enabling reproducible tests and understanding achievable performance on 5G and B5G networks [16]. This requires to measure QoS KPIs related to service latency, stability, and continuity, and define QoE KPIs that quantify service interactivity performance. Within this context, the ITU-T has issued Recommendation G.1051 in 2023, where it provides guidelines for defining traffic patterns, measuring QoS KPIs, and evaluating a QoE KPI, referred to as *i-score*, as a function of the measured QoS KPIs [13].

The ITU-T methodology requires to instantiate a downlink/uplink (DL/UL) data traffic flow between a client-server pair, e.g., over a 5G network. Traffic characteristics should resemble the data exchange from real services; therefore, it is assumed that the client generates packets (the size of which can change over time) at a given rate, and the server reflects each packet with same or different size, depending on the need for emulating DL/UL-symmetric vs. asymmetric flows. The methodology specifies using User Datagram Protocol (UDP) as the transport protocol; moreover, it recommends using Two-Way Active Measurement Protocol (TWAMP) [17] at the higher layer, which enables same-size packet reflection and can be extended for DL/UL-asymmetric traffic.

Once a traffic pattern is defined, the methodology recommends measuring three QoS KPIs for quantifying service latency, stability, and continuity. The first KPI is obtained by measuring the RTT for each packet pair. The second KPI is evaluated by measuring the Packet Delay Variation (PDV) experienced by the packet pairs [18]. The third KPI is evaluated by measuring the Packet Loss Rate (PLR), i.e., the ratio between *disqualified* packets and the total number of client-generated packets. Packets are considered disqualified if not sent or not received during the test duration, or if they are received at the client side but after a service-dependent RTT budget.

On top of the QoS KPIs, the *i-score* can be then evaluated. The *i-score* model is service-agnostic but the parameters used in the model are service-specific, so as to reflect different requirements. When deriving the *i-score* model, it is assumed that the higher the latency, the lower the interactivity, with a logistic function with service-specific parameters representing such relationship. This allows for transforming RTTs from non-disqualified packet into [%] values, the average of which defines the *i-score* RTT-dependent term, $score_{\mathrm{RTT}}$. PDV and PLR KPIs are then included via the $score_{\mathrm{PDV}}$ and $score_{\mathrm{PLR}}$ terms, each obtained from the corresponding QoS KPIs after simple transformations embedding service-specific parameters. Finally, the *i-score* is measured as follows:

$$i\text{-}score = score_{\mathrm{RTT}} \times score_{\mathrm{PDV}} \times score_{\mathrm{PLR}}. \quad (1)$$

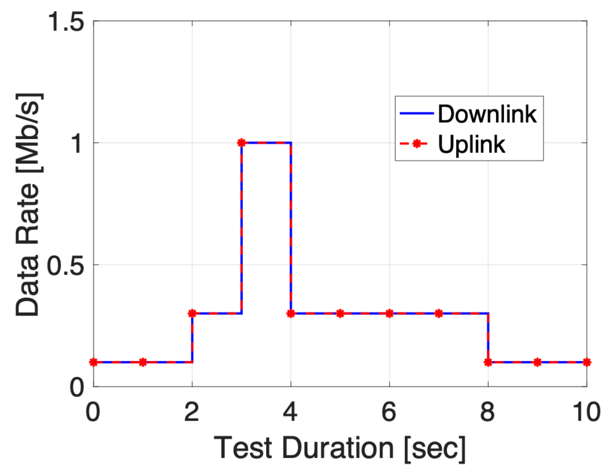


Figure 1: *eGaming real-time* traffic pattern.

We refer to [13], [19]–[21] for additional details on the *i-score* model as a function of RTT, PDV, and PLR KPIs, and on the service-specific parameters.

Figure 1 shows the traffic pattern used for the analyses reported in this paper, referred to as *eGaming real-time*. The pattern is compliant to the ITU-T methodology and emulates a DL/UL-symmetric interaction between a user and a server running an online gaming application, thus representing a relevant example of highly interactive, cloud-based services supported by 5G systems. Phases with different data rates emulate various levels of interaction, under the assumptions that client and server only exchange status information, and heavy video processing is locally performed at the client side. The data rates are obtained by using 100-byte packets, with 125 to 1250 packets generated per second. The RTT budget is $100\,\mathrm{ms}$, following 3GPP specifications [22].

Empirical characterizations leveraging the execution of *eGaming real-time* tests on 5G Non-Standalone (NSA) commercial networks can be found in [23]–[26], which indeed represent some of the first attempts of using the ITU-T methodology for QoS/QoE performance testing on 5G networks.

III. Experimental Campaign

In this paper, we carry out our *i-score* characterisation and prediction by leveraging samples collected while performing a measurement campaign of seven weeks in Rome, Italy. We thus leverage the data open-sourced in [14] in order to a) analyze a specific campaign and characterise the *i-score* trends throughout time and while moving around the city of Rome, and b) apply deeper data-driven analyses that set the basis for our proposed ML-based *i-score* prediction scheme, as further detailed in the following sections.

As detailed in [14], the campaign was conducted with a Samsung S20 5G-capable device. Samples were collected on the 5G NSA networks deployed by two Italian operators (Op1 and Op2) in the mid band (Band n78, 3300-3800 MHz), in different mobility scenarios: Indoor Static (IS), Outdoor Driving (OD), and Outdoor Walking (OW). The campaign was conducted by following the methodology presented in Section II, i.e., by performing interactivity tests with the *eGaming real-time* traffic pattern, thus meeting the requirements of the ITU-T Recommendation G.1051 [13]. We further observe that the collected measurements are suitable for a comparison with the analyses reported in [15], since the same instruments were adopted and the same parameters were measured (i.e., *i-score*, RTT, PDV, and PLR) .

As in [15], in Fig. 2 and in Fig. 3 we analyse the relationship between *i-score* and RTT across all gaming tests, among different scenarios and with/without 5G capabilities enabled.

Figure 2 shows the *i-score* measured among OW, OD, and IS scenarios. We observe that the *i-score* decreases with higher values of RTT, and reaches zero from around $80\,\mathrm{ms}$ in all cases. IS shows a consistent trend with little variability, maintaining high scores for lower RTTs, and reaching zero first. From $45\,\mathrm{ms}$ to $60\,\mathrm{ms}$, all the scenarios exhibit the same trend, while from $30\,\mathrm{ms}$ to $45\,\mathrm{ms}$, OW shows lower *i-score* values, suggesting that, in this case, other factors such as PDV and PLR may have played a more relevant role in lowering the *i-score*.

Figure 3 shows that the *i-score* exhibits similar behaviours with/without 5G capabilities enabled, suggesting that, compared to 4G, 5G NSA infrastructures may not bring clear benefits in terms of QoE for real-time interactive services, which are indeed expected to gain more from 5G Standalone (SA) deployments. Also, for the two mobile operators involved in the campaign, Fig. 4 shows similar *i-score* values, with a slight better performance carried out by Op2 for high *i-score* values.

IV. *i-score* CHARACTERISATION

We provide an additional characterization of the *i-score* by selecting one campaign out of all the campaigns in [14], referred to as *Campaign 6*. *Campaign 6* is well represented by a large number of samples, and the User Equipment (UE) mode was 5G-enabled, meaning the UE was able to take advantage of both 5G and 4G capabilities. Moreover, this campaign was executed on the 5G NSA network of Op1, in the OD scenario, i.e., while driving around the city of Rome. Figure 5 shows the relationship between *i-score* and the number of lost packets (from which PLR can be inferred), PDV, and RTT. Eventually, we also present the throughput observed over both 4G and 5G networks during the campaign (note that, thanks to the 5G NSA infrastructure, UEs can use dual-connectivity

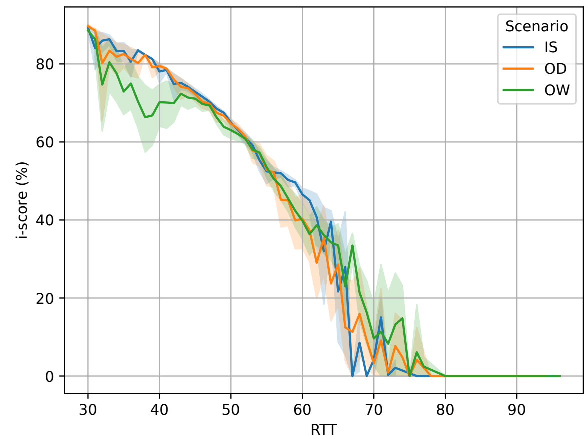


Figure 2: *i-score* [%] across Indoor Static (IS), Outdoor Driving (OD), and Outdoor Walking (OW) scenarios as a function of RTT [ms], for all the campaigns in [14].

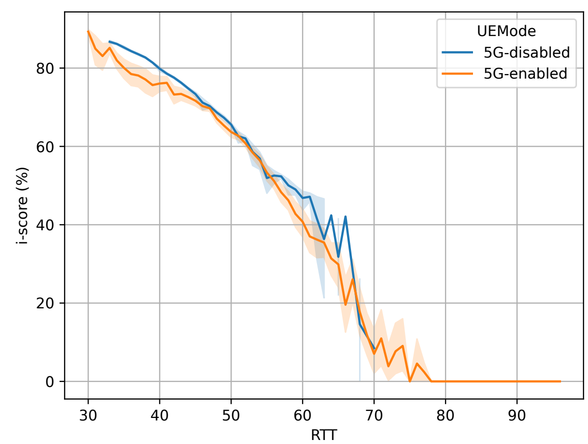


Figure 3: *i-score* [%] in 5G-enabled and 5G-disabled UE Mode as a function of RTT [ms], for all the campaigns in [14].

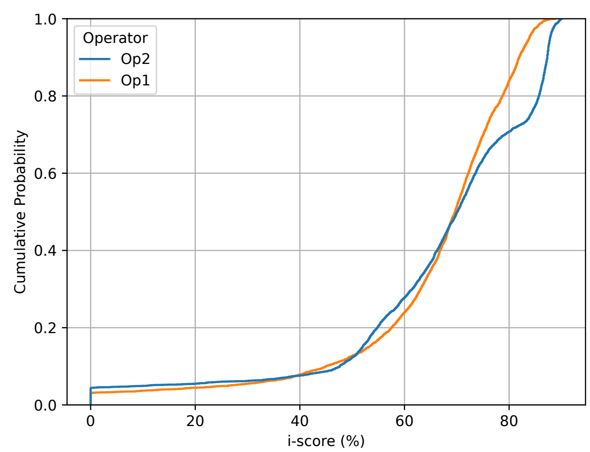
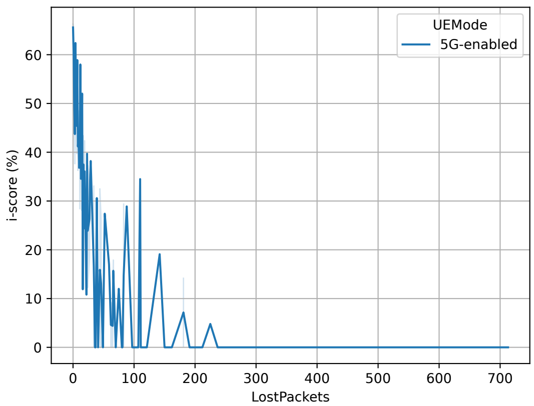
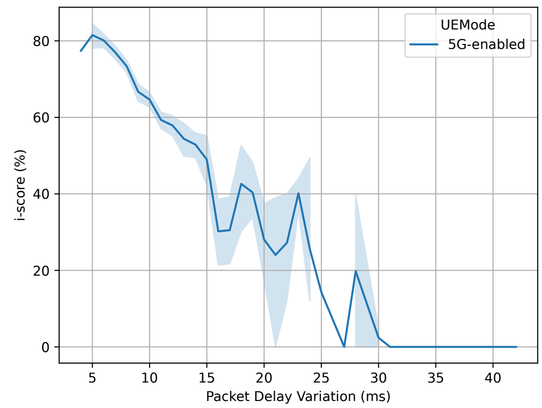


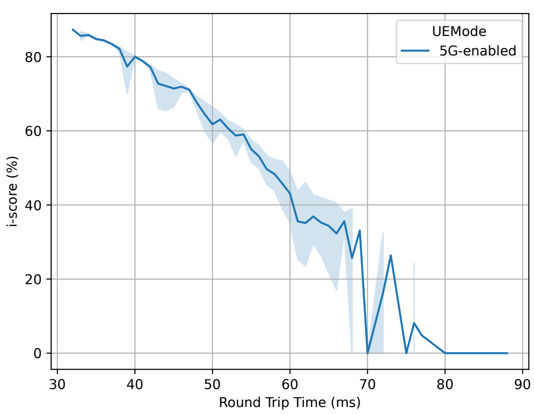
Figure 4: *i-score* ECDF across Operators (Op1 and Op2), for all the campaigns in [14].



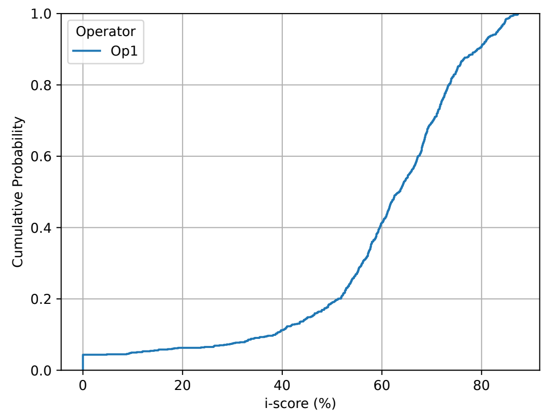
(a) *i-score* vs. Lost Packets.



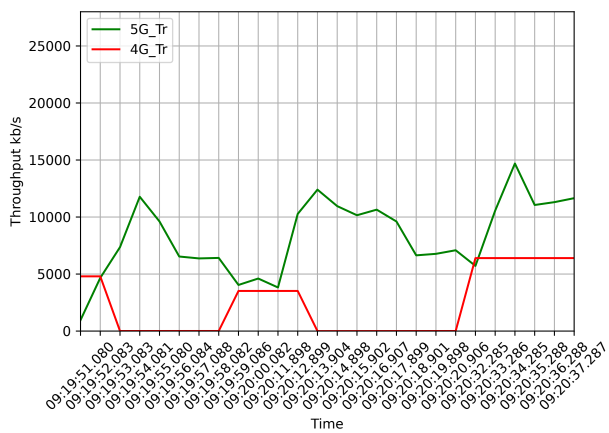
(b) *i-score* vs. PDV.



(c) *i-score* vs. RTT.



(d) *i-score* ECDF.



(e) 4G/5G PHY layer downlink throughput observed during the campaign.

Figure 5: Performance observed during *Campaign 6* (Scenario: OD, UE Mode: 5G-enabled, Operator: Op1).

and exploit 4G/5G networks simultaneously). The aim of this characterisation is to understand whether *i-score* is able to capture the network functioning/not functioning properly in a real scenario. In Fig. 5a, the relationship between *i-score* and the number of Lost Packets is shown. The *i-score* decreases when the number of lost packets increases (when more than 250 packets in a test are lost, the *i-score* is actually zero). Hence, network performance in terms of packet losses are well reflected in the *i-score* values. By comparing Fig. 5b and Fig. 5c with the theoretical *i-score* models reported in

[15], we observe that, as expected, our curves are affected by delay variation and loss encountered on real networks. In particular, Figure 5b represents *i-score* vs. PDV, and shows higher *i-score* values when PDV is under $25\,\mathrm{ms}$. In general, low *i-score* values correspond to high PDV values, which confirms the ability of summarizing network performance in terms of latency variation via the *i-score* parameter. Finally, Fig. 5c presents the variation of *i-score* as a function of RTT, exhibiting a well-expected decreasing trend and thus the ability of reflecting network performance in terms of latency into *i-score* values. To summarize, results in Fig. 5 confirm the ability of the *i-score* to properly describe almost all selected KPIs synthetically.

In [27], good values of *i-score* are considered ranging from 75% and 85% in a 4G network, and from 85% to 90% in a 5G network. For the aforementioned values, a fair to good real-time *eGaming* experience can be met. In our case, as shown in Fig. 5d, the *i-score* for the selected campaign spans from 40% to beyond 80%. It gradually grows up to 40%, increasing rapidly afterwards, with the majority of sample spanning from 40% to 80%. Following the aforementioned ranges, the network under study seems to provide medium-to-good performance.

In Fig. 5e, we observe that the throughput measured during *Campaign 6* on the 4G and 5G Physical (PHY) layer is coherent with 5G NSA architecture, where 5G comes in support of a well established 4G connection. Overall, the relatively low PHY throughput observed during the experiments is limited by the actual application data rate of the traffic pattern (Figure 1); 5G still exhibits a higher PHY throughput (up to $15\,\mathrm{Mbit\,s^{-1}}$) compared to 4G (up to $6\,\mathrm{Mbit\,s^{-1}}$).

V. ML-based *i-score* Prediction

In this section, we test the ability of different supervised ML algorithms in solving a binary classification problem defined on our dataset. In particular, we tried to predict good vs. bad *i-score* values by only looking at radio layer quality and power parameters. Therefore, a threshold was selected and the *i-score* was split among two classes, good values of *i-score* and bad values of *i-score*. The threshold was chosen in order to obtain a balanced dataset and following Fig. 4, and set as the median value, i.e., 70%. Several features from the radio layer were collected during the experimental campaign, as further detailed in the following. For the ML techniques to use, we focused on ensemble methods, considering that, compared to other methods, they provided the best performance in our previous studies, where the goal was to predict RTT in 5G NSA networks [11], [28]–[35].

A. Features and Target

The dataset presents ~ 8400 rows with 12 features:

- 5G Synchronization Signal Reference Signal Received Power (SS-RSRP)
- 5G Synchronization Signal Reference Signal Received Quality (SS-RSRQ)
- 5G Synchronization Signal Reference Signal to Interference plus Noise Ratio (SS-SINR)
- 4G Reference Signal Received Power (RSRP)
- 4G Reference Signal Received Quality (RSRQ)
- 4G Signal-to-Interference-plus-Noise Ratio (SINR)
- 5G PHY layer DL Throughput
- 4G PHY layer DL Throughput
- UEMode 5G-enabled
- Scenario OD
- Scenario OW
- Operator Op2

Table I: P-Value and F-Score.

Feature	P-value	F-score
SS-RSRP	7.4988e-110	462.82
SS-RSRQ	2.4944e-99	167.77
SS-SINR	4.9240e-83	383.24
RSRP	1.0311e-39	139.38
RSRQ	1.0975e-39	176.08
SINR	6.4595e-38	71.349
5G Tr	7.4922e-32	514.70
4G Tr	5.9772e-20	0.0018
UEMode 5G-enabled	3.6332e-17	84.156
Scenario OD	8.9180e-08	28.659
Scenario OW	0.15826	176.21
Operator Op2	0.96615	1.9913

To discover which among these features were able to predict better the *i-score* we carried out a feature importance analysis. As shown in Table I, quality and power parameters as RSRP, RSRQ and SINR for both 4G and 5G play a key role in predicting the target, confirming that it is possible to predict *i-score* performance from those features.

B. Data Processing

As also reported in [14], the collected dataset presents the symbol `?` (question mark) to indicate unaltered values. Indeed, new numerical values appear in the dataset only when actually measured by the system, therefore, `?` indicates that the oldest measured value for that feature is still valid, since a new value has not been recorded yet. Therefore, the `ffill` method[1] was selected to replace all `?` values, with respect of the *Campaign ID*. To strengthen the process, we shuffle the dataset and encoded categorical variables such as Operator, UEMode, and Scenario with Scikit-Learn's One Hot Encoder.[2] Variables were thus encoded by dropping the first column of all new created variables to reduce collinearity. This resulted in 12 original features to be used in our ML-based analysis. Data was also standardized to ensure all algorithms to function smoothly through Scikit-Learn's Standard Scaler[3] before our *i-score* binary classification. Performance statistics, together with all radio layer parameters, were collected with the UE in 5G-enabled mode, which allows for simultaneously using 4G and 5G access networks, as previously discussed. We also filtered all values of RTT equal to $0\,\mathrm{ms}$, since they can be accounted as lost packets or wrong collected measurements.

[1] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.ffill.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Table II: Best parameters for each algorithm.

Algorithm	Best Parameters
Decision Tree	max depth: 10 min samples split: 2 min samples leaf: 1 criterion: gini
Random Forest	num estimators: 300 max depth: 20 min samples split: 2 min samples leaf: 2
Gradient Boosting	num estimators: 300 max depth: 40 min samples split: 10 min samples leaf: 4

C. ML Algorithms and Performance

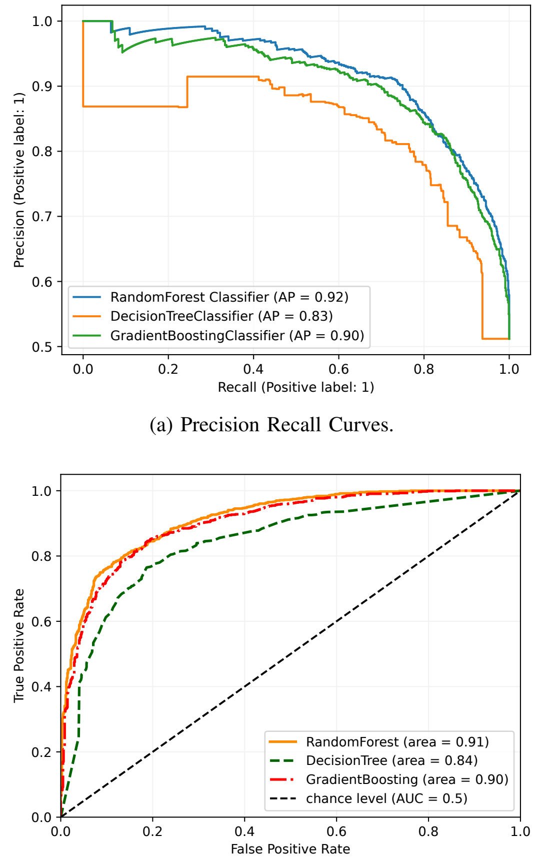
We adopted Decision Tree (DT), Gradient Boosting (GB), and Random Forest (RF) after fine-tuning them through search grid, which resulted in the configurations shown in Table II.

We evaluate the performance of all the aforementioned algorithms in the *i-score* binary classification task, in terms of accuracy, recall and F1-score, as reported in Table III. We observe that GB and RF achieve the same performance, with RF performing slightly better in terms of accuracy for Class 1 (good *i-score* values). DT accuracy is below the other techniques, probably due to the nature of this algorithm, which adopts a single classification tree compared to GB and RF, which instead take advantage of the combination of multiple trees.

To further confirm and analyze the performance of the selected algorithms, we report Precision Recall curves in Fig. 6a, which ultimately show RF and GB as best performing techniques, with an Average Precision (AP) around 0.90. In Fig. 6b, Receiver Operating Characteristic Curves (ROCs) are depicted, where again DT is outperformed by the other two algorithms with an Area Under the Curve (AUC) of 0.84. Eventually, RF results the best performing algorithm for our *i-score* classification. Achieved performances are overall of a good quality, based on the need for less complexity or a higher accuracy, one algorithm can be chosen among the others.

Table III: Precision, Recall, and F1-Score.

Algorithm	Precision	Recall	F1-score	Accuracy	Class
GB	0.81 0.84	0.84 0.81	0.82 0.83	0.83	0 1
RF	0.81 0.84	0.84 0.82	0.83 0.83	0.83	0 1
DT	0.77 0.81	0.81 0.76	0.79 0.79	0.79	0 1



(a) Precision Recall Curves.

(b) Receiver Operating Characteristic Curves.

Figure 6: ML-based *i-score* classification: Performance of RF, GB, and DT algorithms.

VI. Conclusions

This paper provides an empirical analysis and ML-based prediction of QoE on 5G networks, particularly focusing on the *i-score* parameter, a synthetic QoE KPI recently standardized by ITU-T. By exploiting a large-scale measurement campaign carried out in Rome on two 5G NSA deployments, we characterize the observed *i-score*, confirming its ability to synthetically describe QoS KPIs commonly used to understand real-time service performance, i.e., RTT, PDV, and PLR. Additionally, we propose a ML-based approach to predict *i-score* classes based on radio layer parameters. Results confirms the validity of our approach, which can be used to infer QoE performance rather than executing time/resource-consuming tests. In particular, the RF algorithm provides one of the best performance in the classification task, with an accuracy of 0.83, precision recall AP of 0.92, and ROC AUC of 0.91. In future work, we plan to move towards deep learning schemes aiming to further improving the classification accuracy, while also considering the creation of *i-score* models based on radio layer parameters through regression and/or forecasting approaches.

VII. Acknowledgements

This work was partially supported by project cyberHuman, part of the SERICS program (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and by European Union through the ADAPTO project, part of the RESTART program, NextGenerationEU PNRR, CUP E63C2 2002040007, CP PE0000001.

The work by G.C. and A.B. is supported by the Knowledge Foundation of Sweden.

References

[1] Qualcomm, "What is 5g?" 2024, accessed: 2024-08-22. [Online]. Available: https://www.qualcomm.com/5g/what-is-5g#:~:text=5G%20can%20be%20significantly%20faster,traffic%20capacity%20and%20network%20efficiency

[2] Ericsson, "Ericsson mobility report, june 2024," 2024, accessed: 2024-08-22. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2024

[3] 3rd Generation Partnership Project (3GPP), "5g system overview," https://www.3gpp.org/technologies/5g-system-overview, 2024, accessed: 2024-08-12.

[4] International Telecommunication Union, "Itu-t recommendation p.915: Subjective assessment methods for 3d video quality," ITU-T, Standard P.915, 2016, accessed: 2024-08-12. [Online]. Available: https://www.itu.int/rec/T-REC-P.915/en

[5] ——, "Itu-t recommendation p.10: Vocabulary for performance, quality of service, and quality of experience," ITU-T, Standard P.10, 2006, accessed: 2024-08-12. [Online]. Available: https://www.itu.int/rec/T-REC-P.10/en

[6] ——, "Itu-t recommendation p.10 amendment 2: Vocabulary for performance, quality of service, and quality of experience - new definitions for inclusion in recommendation itu-t p.10/g.100," ITU-T, Standard P.10 Amd 2, 2013, accessed: 2024-08-12. [Online]. Available: https://www.itu.int/rec/T-REC-P.10-201311-I!Amd2/en

[7] H.-W. Kao and E. H.-K. Wu, "Qoe sustainability on 5g and beyond 5g networks," *IEEE Wireless Communications*, vol. 30, no. 1, pp. 118–125, 2023.

[8] Y. Zhou, J. Chen, M. Zhang, D. Li, and Y. Gao, "Applications of machine learning for 5g advanced wireless systems," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, 2021, pp. 1700–1704.

[9] G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, "A survey on multimedia services qoe assessment and machine learning-based prediction," *IEEE Access*, vol. 10, pp. 19 507–19 538, 2022.

[10] H.-S. Huang and Y.-S. Su, "A practical study of qoe on cloud gaming in 5g networks," in *2023 International Wireless Communications and Mobile Computing (IWCMC)*, 2023, pp. 638–643.

[11] S. Zinno, A. Botta, G. Ventre, and N. Pasquino, "Prediction of round-trip time in 5g nsa networks in urban environment," in *2024 IEEE International Symposium on Measurements & Networking (M&N)*, 2024, pp. 1–6.

[12] A. Milde and S. Z. Pilinsky, "Comparison of 4g and 5g nr nsa qoe measurements in croatian cities," in *2022 International Symposium ELMAR*, 2022, pp. 13–18.

[13] ITU-T, "Latency measurement and interactivity scoring under real application data traffic patterns," *Recommendation ITU-T G.1051 (2023-03)*, 2023.

[14] K. Kousias, M. Rajiullah, G. Caso, U. Ali, O. Alay, A. Brunstrom, L. De Nardis, M. Neri, and M.-G. Di Benedetto, "A Large-Scale Dataset of 4G, NB-IoT, and 5G Non-Standalone Network Measurements," *IEEE Communications Magazine*, vol. 62, no. 5, pp. 44–49, 2024.

[15] J. Sochos. (2020, April) Interactivity test: Concept and kpis (part 2). Accessed: 2024-07-22. [Online]. Available: https://www.rohde-schwarz.com/us/solutions/test-and-measurement/mobile-network-testing/stories-insights/article-interactivity-test-concept-and-kpis-part-2-_253245.html

[16] A. Díaz Zayas, G. Caso, Ö. Alay, P. Merino, A. Brunstrom, D. Tsolkas, and H. Koumaras, "A modular experimentation methodology for 5g deployments: The 5genesis approach," *Sensors*, vol. 20, no. 22, p. 6652, 2020.

[17] IETF, "A two-way active measurement protocol (twamp)," *IETF RFC 5357*, 2008, accessed on August 2024. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5357

[18] ——, "Packet Delay Variation Applicability Statement," *IETF RFC 5481*, 2009, accessed on August 2024. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5481

[19] ETSI, "Speech and multimedia transmission quality (stq); qos parameters and test scenarios for assessing network capabilities in 5g performance measurements," *ETSI TR 103 702 V1.1.1 (2020-11)*, 2020.

[20] ——, "Speech and multimedia transmission quality (stq); best practices for robust network qos benchmark testing and scoring," *ETSI TR 103 559 V1.2.1 (2023-10)*, 2023.

[21] R. Schwarz, "Interactivity test: Qos/qoe measurements in 5g (part 1)," 2020, accessed on August 2024. [Online]. Available: https://www.rohde-schwarz.com/us/solutions/test-and-measurement/mobile-network-testing/stories-insights/article-interactivity-test-qos-qoe-measurements-in-5g-part-1-_253244.html

[22] 3GPP, "5G; System Architecture for the 5G System (Rel-15)," Technical Specification (TS) 23.501, 2018.

[23] K. Kousias, M. Rajiullah, G. Caso, O. Alay, A. Brunstrom, L. De Nardis, M. Neri, U. Ali, and M.-G. Di Benedetto, "Coverage and performance analysis of 5g non-standalone deployments," in *Proceedings of the ACM MobiCom Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, 2022.

[24] ——, "Implications of handover events in commercial 5g non-standalone deployments in rome," in *Proceedings of the ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases*, 2022, pp. 22–27.

[25] G. Caso, M. Rajiullah, K. Kousias, U. Ali, L. De Nardis, A. Brunstrom, O. Alay, M. Neri, and M.-G. Di Benedetto, "An initial look into the performance evolution of 5g non-standalone networks," in *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2023, pp. 1–4.

[26] K. Kousias, M. Rajiullah, G. Caso, O. Alay, A. Brunstrom, U. Ali, L. De Nardis, M. Neri, and M.-G. Di Benedetto, "Empirical performance analysis and ml-based modeling of 5g non-standalone networks," *Computer Networks*, vol. 241, p. 110207, 2024.

[27] J. Sochos. (2020, April) Interactivity test: Examples from real 5g networks (part 3). Accessed: 2024-07-22. [Online]. Available: https://www.rohde-schwarz.com/us/solutions/critical-infrastructure/mobile-network-testing/stories-insights/article-interactivity-test-examples-from-real-5g-networks-part-3-_253380.html

[28] S. Avallone, N. Pasquino, S. Zinno, and D. Casillo, "Experimental characterization of LTE adaptive modulation and coding scheme under actual operating conditions," in *2017 IEEE International Workshop on Measurement and Networking, M and N 2017 - Proceedings*, 2017, pp. 1–6.

[29] N. Pasquino, G. Ventre, S. Zinno, F. Ignarro, and S. Petrocelli, "Multivariate LTE Performance Assessment through an Expectation-Maximization Algorithm Approach," in *2019 IEEE International Symposium on Measurements and Networking, M and N 2019 - Proceedings*, 2019, pp. 1–5.

[30] S. Avallone, N. Pasquino, G. Ventre, and S. Zinno, "Experimental Characterization of Long Term Evolution Multiple Input Multiple Output Performance in Urban Propagation Scenarios," in *2018 Workshop on Metrology for Industry 4.0 and IoT, MetroInd 4.0 and IoT 2018 - Proceedings*, 2018, pp. 254–258.

[31] ——, "Smartphone-based drive-test methodology for the experimental characterisation of the physical layer and a performance assessment of LTE networks in urban environments," *Acta IMEKO*, vol. 8, no. 2, pp. 28–34, 2019.

[32] N. Pasquino, S. Zinno, F. Cotugno, and S. Petrocelli, "A comparative approach of unsupervised machine learning techniques for LTE network parameter clustering," in *I2MTC 2020 - International Instrumentation and Measurement Technology Conference, Proceedings*, 2020, pp. 1–6.

[33] N. Pasquino, G. Ventre, S. Zinno, and S. Petrocelli, "Multivariate Analysis of LTE Radio-Layer Parameters based on a Partitional Clustering Approach," in *2019 IEEE International Workshop on Metrology for Industry 4.0 and IoT, MetroInd 4.0 and IoT 2019 - Proceedings*, 2019, pp. 22–27.

[34] S. Avallone, N. Pasquino, S. Zinno, and D. Casillo, "Smartphone-based measurements of LTE network performance," in *I2MTC 2017 - 2017 IEEE International Instrumentation and Measurement Technology Conference, Proceedings*, 2017, pp. 1–6.

[35] S. Zinno, A. Affinito, N. Pasquino, G. Ventre, and A. Botta, "Prediction of RTT Through Radio-Layer Parameters in 4G/5G Dual-Connectivity Mobile Networks," in *2023 IEEE Symposium on Computers and Communications (ISCC)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2023, pp. 213–218.