A No-Reference Video Streaming QoE Estimator based on Physical Layer 4G Radio Measurements

D. Moura[1,4] and M. Sousa[1,2,4]
[1]Instituto Superior Técnico
[2]CELFINET,
Consultoria em Telecomunicações, Lda.
Lisbon, Portugal
diogofmmoura@tecnico.ulisboa.pt,
marco.sousa@celfinet.com

P. Vieira[3,4]
[3]Instituto Superior de Engenharia Lisboa
Lisbon, Portugal
pedro.vieira@isel.pt

A. Rodrigues[1,4] and M. P. Queluz[1,4]
[4]Instituto de Telecomunicações,
Lisbon, Portugal
[ar, paula.queluz]@lx.it.pt

***Abstract*—With the increase in consumption of multimedia content through mobile devices (e.g., smartphones), it is crucial to find new ways of optimizing current and future wireless networks and to continuously give users a better Quality of Experience (QoE) when accessing that content. To achieve this goal, it is necessary to provide Mobile Network Operator (MNO) with real time QoE monitoring for multimedia services (e.g., video streaming, web browsing), enabling a fast network optimization and an effective resource management. This paper proposes a new QoE prediction model for video streaming services over 4G networks, using layer 1 (*i.e.*, Physical Layer) key performance indicators (KPIs). The model estimates the service Mean Opinion Score (MOS) based on a Machine Learning (ML) algorithm, and using real MNO drive test (DT) data, where both application layer and layer 1 metrics are available. From the several considered ML algorithms, the Gradient Tree Boosting (GTB) showed the best performance, achieving a Pearson correlation of 78.9%, a Spearman correlation of 66.8% and a Mean Squared Error (MSE) of 0.114, on a test set with 901 examples. Finally, the proposed model was tested with new DT data together with the network's configuration. With the use case results, QoE predictions were analyzed according to the context in which the session was established, the radio transmission environment and radio channel quality indicators.**

***Index Terms*—Mobile Wireless Networks, LTE, Video Streaming, Quality of Experience, Machine Learning.**

I. Introduction

One of the most prominent issues that current and next-generation wireless network operators will face is assuring high Quality of Experience (QoE) to an increasing number of users, in services that generate a high network load, as video streaming [1]. In this context, much work has been done on assessing Quality of Service (QoS) through the measurement of specific network key performance indicators (KPIs) and drive tests (DTs) (e.g., [2] [3] [4]). However, these metrics cannot directly assess the user's QoE. Most of the proposed video streaming QoE models (e.g., [5] [6]) rely on application layer metrics that are not measurable in real-time by the Mobile Network Operators (MNOs) being only available at the client side; thus, these models are inadequate to predict the QoE from the MNO's perspective. Hence, MNOs have limited capability to obtain a proper gauge of the network's QoE status at any given time, undermining near real time monitoring and optimization.

Some work has been developed regarding the QoE estimation for other services, namely voice and web browsing [7], using real network data. Concerning video streaming QoE estimation, although it has been an active research area, to the best of the authors knowledge, most of the works target analytical and simulation approaches (e.g., [8] [9]), lacking contributions with real MNO data. In this work, a QoE model for predicting the quality that a user experiences in a video streaming session through a 4G network, based on objective radio channel QoS metrics is proposed. To this end, Machine Learning (ML) techniques are used with the goal of building a mathematical model between network QoS metrics and a QoE metric, the Mean Opinion Score (MOS), based on data obtained from several dedicated DTs in a live 4G network.

The paper is organized as follows: in Section II the used data is analyzed; Section III presents the QoE model development process and results; Section IV presents a real scenario where the model is applied. Finally, conclusions are drawn in Section V.

II. Data Analysis

The development of the proposed QoE model was supported by data provided from dedicated DT campaigns, where video streaming services were established and monitored with the following conditions:

- The video streaming service was tested by playing a Youtube video, from a set of three different videos, two of them lasting 45 s and one lasting 47 s.
- Several application layer metrics were measured during the Youtube video session (*e.g.*, number of video freezes, average video resolution) and with the layer 1 protocol (*e.g.*, Channel Quality Indicator (CQI), Reference Signal Received Quality (RSRQ), Scheduled Throughput).
- Three MNOs of 4G networks were tested.
- Two distinct mobile devices, Samsung Galaxy S8 (SM-G950F) and Sony XZ (F8331) were used.

978-1-7281-3106-1/20/$31.00 ©2020 IEEE

- The DT campaigns were conducted in different locations (cities, highways, suburban and others) and mobility settings (no mobility or at different speeds);

From the DT campaigns, the sessions with missing KPIs or with incomplete data due to failure in initiating the video stream were removed. This resulted in a dataset for the model development with 4510 video streaming sessions of which 80% (3608 sessions) were used as training data and the remaining sessions (901 sessions) were used for testing the proposed model performance. For each session, 60 network KPIs were retained together with the available application layer metrics.

A. QoE Reference Model

In adaptive video streaming, the server has different representations of the same video, each one with a different bitrate (and subjective quality), and each representation is divided into segments, which can be independently decoded by the client. During the streaming session, the client may switch between the different video representations, in order to cope with channel throughput fluctuations. The main application layer metrics that condition the QoE along the streaming session are [10]: initial playout delay (due to initial buffering at the client side), average quality of the transmitted video, frequency and amplitude of the video quality switches, video freezes (due to empty buffer) frequency and average video freeze duration. The objective model proposed in [5] incorporates all of these metrics, and was used in the present work to obtain the MOS for each streaming session (and since the subjective assessments, obtained with real users, were not available); it is formally described by:

$$MOS_{est} = 5.67 * \frac{\bar{q}}{q_{max}} - 6.72 * \frac{\hat{q}}{q_{max}} - 4.95 * F + 0.17 \quad (1)$$

where $\bar{q}$ represents the average video quality level (requested by the client) and $\hat{q}$ as its standard deviation; both are normalized with respect to the highest available quality level, q_{max}, for that video. Parameter F models the influence of freezes and is given by:

$$F = \frac{7}{8} * \max\left(\frac{\ln(\phi)}{6} + 1, 0\right) + \frac{1}{8} * \left(\frac{\min(\psi, 15)}{15}\right) \quad (2)$$

where ϕ and ψ represents the freeze frequency and the average freeze duration, respectively. In this work, since Youtube generates the different representations for the same video by changing their spatial resolution, the $\bar{q}$ and q_{max} parameters are given by the average transmitted video spatial resolution and by the maximum video spatial resolution available for that video, respectively, which are directly available on the dataset. Both ϕ and ψ were obtained through simple calculations using some of the parameters provided in the dataset, as presented:

$$\phi = \frac{Interruptions\,[\#]}{Duration\,[s]} \quad (3)$$

$$\psi = \frac{Interruptions\,[s]}{Interruptions\,[\#]} \quad (4)$$

The $\hat{q}$ parameter could not be obtained since the data regarding the changes in segment quality required for this calculation was not available in the dataset. The $\hat{q}$ was set to zero assuming that due to the small size of smartphone screens in which these videos were/are going to be watched, the changes in resolution does not greatly impact the users' QoE. After computing the parameters of (1) for each session, their corresponding MOS values were obtained. Since (1) results in MOS values in the range of [0, 5.84], these were linearly rescaled to the usual range, [1 5]. The rescaled MOS values are shown in Figure 1.

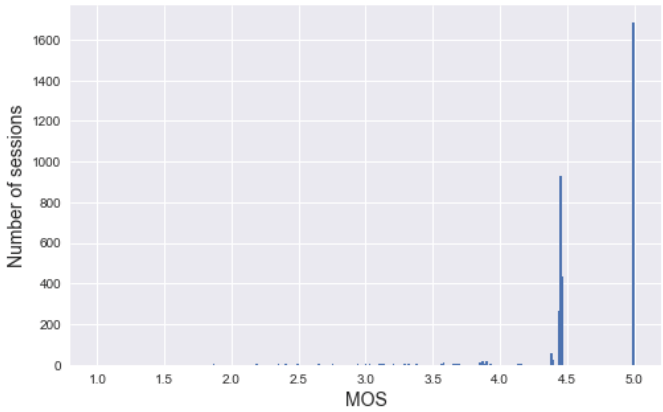

Fig. 1: MOS distribution (all sessions).

From Figure 1, it can be seen that the dataset is extremely unbalanced, where the MOS values in the interval]4, 5] are at a clear majority. This may result in a possible difficulty for the ML process to accurately predict the less represented, lower values of MOS, which is not desirable since these are the sessions in which the user is experiencing critical levels of video streaming quality to which the MNO needs to be alerted to.

B. Dimensionality Reduction

Since 60 network KPIs per streaming session were available for model development, a feature selection process was carried out with the purpose of reducing the ML training process complexity and overfitting. In the first stage, and to eliminate only features providing redundant information, the Pearson Correlation Coefficient (PCC) between all KPIs was computed; the resulting correlation matrix is presented in Figure 2.

To eliminate the most correlated features, a threshold of 0.8 was applied to the PCC values, reducing the number of KPI features to 37.

In the second correlation stage, PCC and Spearman Correlation Coefficient (SCC) between the remaining features and the respective MOS values were computed. After obtaining the Cumulative Distribution Function (CDF) of the PCC and SCC values, it was verified that the 50th percentiles were, respectively, 0.143 and 0.089. Given these low correlation values, the KPIs included in the lower 50th percentile on both

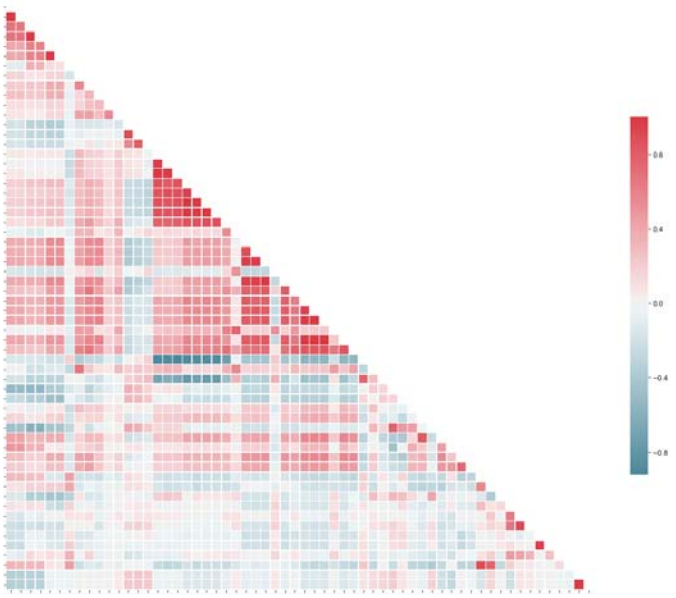

Fig. 2: Heatmap of the Pearson correlation matrix between all 60 network KPIs.

CDF distributions were discarded, resulting in 22 KPI features to be used on the model development.

III. QoE Model Development

The proposed model aims to predict the MOS values of video streaming sessions, from the appropriate network KPIs - available in real time to the MNOs - and was obtained using a learning algorithm over DT data provided by MNOs. According to [11], ensemble methods - where multiple learning algorithms are combined (e.g., decision trees) - may offer significant improvements in both robustness to skewed distributions and in predictive power, when used with unbalanced training sets, which is the case in this work.

For this reason, several ensemble algorithms were considered during model development, namely: Gradient Tree Boosting (GTB), Extra Trees (ET), Random Forest (RF) and Adaptive Boosting (AB). Additionally, the Support Vector Regression (SVR) algorithm was used as a comparison between ensemble-based methods and a commonly used regression algorithm.

A. Performance Metrics

During model development, the following performance metrics were used:

- **Correlations** - Spearman and Pearson correlations were used to assess the relationship between predicted and ground truth MOS values;
- **10-Fold Cross Validation (CV) using Mean Squared Error (MSE)** - Allows for tuning the ML model hyperparameters by splitting the training data into ten subsets of data, where in each iteration nine subsets are used for training and remaining one is used for validation. The best hyperparameter configuration is the one that results in the lowest MSE between predicted and ground truth MOS values;
- **Stratified Error** - By splitting the dataset in four MOS strata (i.e., [1, 2],]2, 3],]3, 4],]4, 5]), the MSE can be computed per stratum, allowing a higher granularity for the error analysis. In fact, due to the skewed MOS distribution towards the highers values, if the MSE is measured on the whole MOS range it will be mainly influenced by the error in the higher MOS values, preventing a proper assessment of the model performance in the lower MOS range.
- **Mean Absolute Scaled Error (MASE)** - The MASE compares the Mean Absolute Error (MAE) of the model's predictions with the predictions of a naïve model [12], which, for this dataset, can be a model that simply outputs the median of MOS values;
- **R-Squared** - Represents the proportion of variance of the prediction results that have been explained by the independent variables (features) of the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model.

B. Data Balancing and ML algorithms Analysis

In order to address the concerns posed at the end of Section II-A regarding the possible learning impairments when the ML algorithms receive as input a severely unbalanced dataset, the initial training set was modified so that all MOS strata were more evenly represented in the resulting distribution. To test the hypothesis that the prediction accuracy of the more critical values of MOS - those belonging to the strata [1, 2],]2, 3] - could increase, 90% of the sessions belonging to the interval]4,5] were removed, at random, from the training set, in order to achieve a more balanced training set while still capturing the tendency, from the original distribution, of the highest MOS interval of being the most represented. Ten models were then created using the unbalanced and balanced training sets, and the five aforementioned ML algorithms; the models prediction performance was analyzed using the metrics listed in Section III-A, and the results are presented in Table I. In this table, the darker colored entries identify, for each performance metric and algorithm, the training set with best performance, balanced or unbalanced. From the results' analysis the following can be stated:

- For all ML algorithms, the majority of the performance metrics indicates that the unbalanced training set produces the best results; however, from the Stratified Error metric, it is visible that there is a decrease in the MSE from the unbalanced to the balanced training sets, in the intervals of]2, 3] and]3, 4], and only a marginal increase in the interval [1, 2] for ET and AB algorithms. Additionally, an increase in MSE is verified from the unbalanced to the balanced training sets in the]4,5] stratum which is not significant since it is more important to accurately predict the MOS in the lower strata. The decrease in MSE, seen in the intervals]2, 3],]3, 4]

TABLE I: Balanced (B) and Unbalanced (UB) training set performance comparison.

Performance Metrics		GTB (UB)	GTB (B)	ET (UB)	ET (B)	RF (UB)	RF (B)	AB (UB)	AB (B)	SVR (UB)	SVR (B)
Test Set Correlations	Pearson	0.897	0.795	0.890	0.798	0.891	0.724	0.896	0.714	0.848	0.703
	Spearman	0.826	0.679	0.802	0.692	0.831	0.647	0.814	0.629	0.777	0.611
Test Set Error	MSE	0.040	0.109	0.044	0.103	0.043	0.153	0.041	0.158	0.058	0.166
	R^2	0.805	0.471	0.785	0.503	0.794	0.261	0.803	0.238	0.711	0.195
Training Set Error	10-Fold CV	0.043	0.201	0.049	0.188	0.043	0.209	0.043	0.212	0.372	0.504
Test Set Stratified Error	1-2	0.451	0.318	0.740	0.818	1.039	0.385	0.473	0.501	1.482	0.205
	2-3	0.990	0.477	1.495	0.604	0.756	0.432	1.241	0.536	0.923	0.508
	3-4	0.504	0.210	0.459	0.220	0.445	0.254	0.516	0.308	0.412	0.303
	4-5	0.009	0.100	0.007	0.089	0.014	0.144	0.006	0.146	0.029	0.157
MASE		0.204	0.652	0.224	0.609	0.197	0.647	0.188	0.714	0.401	0.911

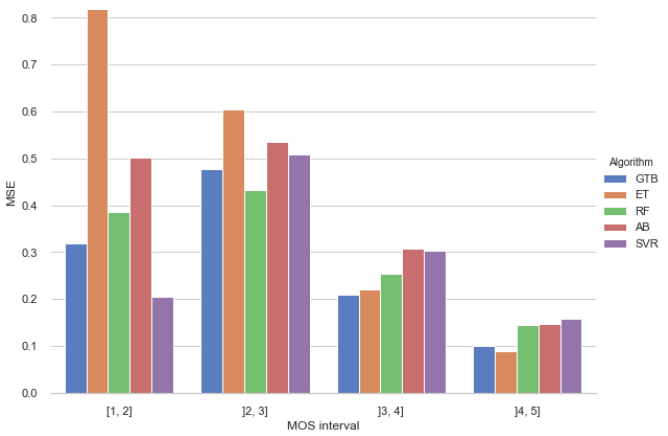

Fig. 3: Barplot of Stratified Error results (all algorithms).

for all algorithms and [1, 2] for the GTB, RF and SVR algorithms, proves that when a ML algorithm uses a balanced training set, it can produce more accurate predictions for less represented values than when using an unbalanced training set;

- All considered ML algorithms could be used, with varying degrees of performance and complexity, to obtain acceptable MOS predictions;
- Both the GTB and RF have similar performances when using the balanced training set in terms of the Stratified Error, and are able to outperform the ET and AB algorithms in this particular metric for the two lowest intervals of MOS. Furthermore, the GTB algorithm outperforms the RF in eight of the performance metrics, being only marginally worse in two metrics.
- The SVR produces the best predictions for most of the lower stratified error intervals but performs significantly worse than the other algorithms in all other metrics.

The correct predictions of the lowest MOS interval is the most important since these values correspond to the most critical QoE that a user has when watching a video; such poor experience, when accurately predicted, can alarm a MNO allowing for improvements to be made to its network, ultimately increasing customer satisfaction and decreasing service churn rates. Thus, the balanced training set was used to train the final model.

Table I shows that the best performing ML algorithm in most of the metrics is the ET algorithm. However, in what concerns the Stratified Error, and as can be seen in Figure 3, the ET algorithm underperforms the others in the two lowest intervals of MOS values, which is not desirable at all. As such, the algorithm that ensues is the GTB, which has the second best performance in most metrics and outperforms the ET algorithm in the three lowest intervals of MOS for the Stratified Error. As such, the algorithm that ensues is the GTB, which has the second-best performance in most performance metrics and manages to outperform the ET algorithm in the three lowest intervals of MOS for the stratified error. Together with its low complexity, this makes it a candidate model to be used, as it balances lower and higher MOS predictions accuracy. Therefore, the GTB algorithm model was selected for the QoE model development, as it is expected to produce accurate MOS predictions.

C. Proposed QoE Model

After using the GTB algorithm, the "importance" of each feature was obtained, according to [13]; it indicates the feature relevance for the model learning: the higher the feature importance, the greater the model dependence on that feature. To determine how many non-relevant features could be removed while maintaining the model performance, the features were ranked according to their importance, and several models were trained with an increasing amount of features, starting with the most important ones (*e.g.* the 2nd model was trained with the two most important features while the 10th model was trained with the 10 most important features). The metrics used to analyze each model were the MSE, the 10-Fold CV and the Stratified Error and additionally the Adjusted R-Squared was used. By using the Adjusted R-squared, it can then be assumed that adding the remaining features is not necessary and that these can be discarded, decreasing the model complexity. The results of the model performance with an increasing number of used features is presented in Figure 4.

From this figure, it can be observed that, after the addition of 11 features, most error metrics do not decrease significantly. Therefore, the number of features was set at 11, since the addition of more (and less important) features will not increase the performance of the model, but increase its complexity and the possibility of overfitting. The 11 selected features were separated in three groups:

- **Transmission Rate (80.9% importance)** : Maximum Aggregate Scheduled Throughput, Aggregate Average Scheduled Throughput, Scheduled Throughput;
- **Channel Quality (11.5% importance)** : Uplink (UL) Acknowledgments (ACKs), Reference Signal Received Quality, Downlink (DL) Number of Transport Blocks

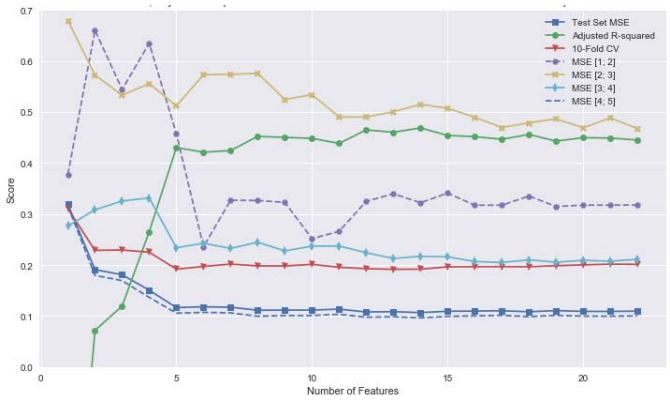

Fig. 4: Model performance with a cumulative number of used features.

(TBs), DL Average Block Error Rate (BLER) and DL Standard Deviation CQI;

- **Modulation and Coding (5.8% importance)** : DL Modulation (MOD) Quadrature Phase Shift Keying (QPSK), DL MOD 64Quadrature Amplitude Modulation (QAM), DL MOD 16QAM.

Regarding the modulation and coding features group, these express the video time percentage using each modulation type, respectively.

The results obtained with the reduced features model on the test set are shown in Table II.

TABLE II: QoE model performance using the GTB algorithm and the reduced set of features.

Correlation		Test Set Error		Training Set Error	Stratified Error				MASE
Pearson	*Spearman*	*MSE*	R^2	*10-Fold CV*	*1-2*	*2-3*	*3-4*	*4-5*	0.660
0.789	0.668	0.114	0.451	0.196	0.266	0.490	0.237	0.103	

From Table II the following comments can be drawn:

- The value for the MSE of the entire test set (i.e., 0.114) is close to the stratified MSE for the interval]4, 5] of the test set (i.e., 0.103), which is expected since this is the interval where most of the ground truth MOS values belong to - even after training set balancing, meaning that the error for the remaining intervals in stratified error do not significantly influence this error. This shows that the the Test Set MSE does not properly gauge the prediction accuracy for the lower MOS values.
- From the Stratified Error metric, it can be noted that the proposed model predicts the lowest intervals of the MOS scale with low error.
- When comparing the 10-Fold CV metric with the MSE for this test set, it can be asserted that the performance verified for this particular test set is similar to that of the entire dataset using CV. This means that the model performance is maintained over several test sets and surely will be maintained when the model is used on new data.

IV. Use Case

After obtaining a new, unseen dataset and with the coordinates of these sessions, a geographical representation of the MOS predictions was obtained, from where it was possible to identify areas with low MOS predictions. These areas were scrutinized by analyzing the radio context in which these sessions took place. In this dataset the information regarding which cell the User Equipment (UE) was connected to, for the duration of the session, was also available. Together with the network configuration information a detailed analysis was conducted. This way, it was possible to conjecture about why the MOS degradation was registered, taking into account the 11 considered radio KPIs and the network configuration leading, to the identification of, for example, poor coverage, interference issues or low capacity.

In Figure 5, an example of a low MOS area, is shown.

Fig. 5: Example area where a prediction of poor MOS is presented.

In it, it is possible to see the path of the vehicle in which these sessions were conducted and their corresponding MOS predictions. Several video sessions present a good (4) or excellent (5) MOS value. However, one session presents a MOS of 3, which could mean that this is an area with poor radio channel conditions. Additionally, the sites to which the UE connected to are Site A and Site B (shown with blue markers in Figure 5). From the dataset, it was possible to verify that another two eNodeBs, positioned further away, also connected to the UE throughout the streaming session. It was also verified, that the RSRQ was low for that particular session, which is probably due to the fact that during the video session the received power from neighboring cells lead to the degradation of the radio channel's conditions which caused a decreased signal to interference plus noise ratio. This resulted in the UE switching to a more robust modulation scheme, in this case QPSK - as was verified in the DT data. A consequence of using a more robust modulation scheme is a decrease in throughput, which has a direct impact in the QoE, since adaptive video streaming is very sensitive to changes in throughput. This variation in throughput leads to a degradation

in quality of adaptive streaming which is correctly pointed out by the proposed QoE prediction model.

V. Conclusions

This paper presents a novel QoE prediction model, for video streaming based on real 4G data that estimates the MOS given the required layer 1 KPIs. From the developed work, it was found that the ML regression process, and consequentially the models' predictions, do benefit from a balancing of the distribution of the training data. For the purposes of substantiating the benefits of training set balancing, there was a need for using different performance metrics, namely the Stratified Error, which measures the MSE for certain MOS intervals. Additionally, 5 different ML algorithms were used and their performances compared, being the GTB algorithm the one that managed to achieve the best results. The model estimates QoE with a Pearson correlation of 78.9%, a Spearman correlation of 66.8% and a MSE of 0.114. Another conclusion is that the features related with the transmission rate of the streaming session are the most important features in predicting QoE for the GTB algorithm. These features assumed a combined feature importance of approximately 81%. Overall, the performed predictions made using layer 1 metrics provide accurate results.

Acknowledgment

The authors would like to thank FCT for the support by the project UID/EEA/50008/2013. Moreover, our acknowledgement concerning project MESMOQoE (n° 023110 - 16/SI/2016) supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

References

[1] Ericsson, "ERICSSON MOBILITY REPORT," Ericsson, Tech. Rep., June 2019.

[2] M. Sousa, A. Martins, and P. Vieira, "Self-Diagnosing Low Coverage and High Interference in 3G/4G Radio Access Networks Based on Automatic RF Measurement Extraction," in *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications*, ser. ICETE 2016, Portugal, 2016, pp. 31–39.

[3] A. Rufini, A. Neri, F. Flaviano, and M. Baldi, "Evaluation of the impact of mobility on typical kpis used for the assessment of qos in mobile networks: An analysis based on drive-test measurements," in *2014 16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, Sep. 2014, pp. 1–5.

[4] R. Santos, M. Sousa, P. Vieira, M. P. Queluz, and A. Rodrigues, "An Unsupervised Learning Approach for Performance and Configuration Optimization of 4G Networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2019, pp. 1–6.

[5] J. De Vriendt, D. De Vleeschauwer, and D. C. Robinson, "QoE model for video delivered over an LTE network using HTTP adaptive streaming," *Bell Labs Technical Journal*, vol. 18, no. 4, pp. 45–62, March 2014.

[6] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for dash video streaming," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 651–665, Dec 2015.

[7] V. Pedras, M. Sousa, P. Vieira, M. P. Queluz, and A. Rodrigues, "A no-reference user centric QoE model for voice and web browsing based on 3G/4G radio measurements," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[8] P. G. Balis and M. R. Tanhatalab, "Analytic Model for the Prediction of Cell-Edge QoE of Streaming Video Over Best-Effort Mobile Radio Bearers," in *2018 26th Telecommunications Forum (TELFOR)*, Nov 2018, pp. 1–4.

[9] Z. Cheng, L. Ding, W. Huang, F. Yang, and L. Qian, "A unified QoE prediction framework for HEVC encoded video streaming over wireless networks," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, June 2017, pp. 1–6.

[10] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for dash video streaming," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 651–665, Dec 2015.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, pp. 221–232, 2016.

[12] R. Hyndman and A. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, pp. 679–688, 02 2006.

[13] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.