Explainable data-driven QoE modelling with XAI

Nikolas Wehner, Anika Seufert, Tobias Hoßfeld, Michael Seufert

Angaben zur Veröffentlichung / Publication details:

Wehner, Nikolas, Anika Seufert, Tobias Hoßfeld, and Michael Seufert. 2023. "Explainable data-driven QoE modelling with XAI." In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX), 20-22 June 2023, Ghent, Belgium*, edited by Hantao Liu, Sam Van Damme, and Tim Wauters, 7–12. Piscataway, NJ: IEEE. https://doi.org/10.1109/qomex58391.2023.10178499.

Explainable Data-Driven QoE Modelling with XAI

Nikolas Wehner, Anika Seufert, Tobias Hoßfeld, Michael Seufert
University of Würzburg, Institute of Computer Science, Würzburg, Germany
{nikolas.wehner | anika.seufert | tobias.hossfeld | michael.seufert}@uni-wuerzburg.de

***Abstract*—Data-driven QoE modelling using Machine Learning (ML) allows to reduce the modelling bias and to continuously integrate new QoE results into the QoE model, which can improve its generalizability. The downside is that the majority of ML models are black-box models, which prevent to obtain insights about QoE influence factors and their fundamental relationships that are highly relevant for researchers and providers of services and networks. However, recent advances in the field of eXplainable Artificial Intelligence (XAI) resolve these issues. Thus, XAI allows to benefit from data-driven QoE modelling to obtain generalizable QoE models, and at the same time to understand what QoE factors are relevant and how they affect the QoE score. In this work, we showcase the feasibility of explainable data-driven QoE modelling for video streaming, since video streaming QoE has been well researched, and thus, allows us to validate our results. Finally, we discuss opportunities and challenges of deploying XAI for QoE modelling.**

I. Introduction

To improve services and networks and to avoid user churn and subsequent revenue losses, researchers and providers of services and networks require a thorough understanding of the factors influencing Quality of Experience (QoE) [1]. To successfully develop a QoE model, dedicated, extensive, and expensive studies are required, which typically can only cover a subset of the parameter space and are influenced by the study design. They often output a relatively small sample of QoE ratings from a comparatively small population, thus being susceptible for poor performance on unseen data. Processing the collected data and performing the actual QoE modelling is not only cumbersome and time-consuming, but might also introduce biases and self-fulfilling prophecies, e.g., seeing an exponential relationship when expecting one.

To bypass these burdens, data-driven QoE modelling with machine learning (ML) is an interesting alternative, in particular, in scenarios where lots of data are available or where streams of data can be obtained continuously. A prime example, which mixes both worlds, is the ITU-T standard P.1203 [2] for estimating video streaming QoE. Its output factors in both manual modelling, accounting for 75% of the estimated mean opinion score (MOS), and Random Forest based ML modelling, accounting for the remaining 25%. Apparently, the ML component improves the performance of P.1203, otherwise, it would not have been included. However, the internals of P.1203's Random Forest model, i.e., how exactly its output score is inferred, are not obvious. This black-box characteristic also holds for most other ML models, meaning that their decision-making processes are generally difficult to understand for humans. This prevents to obtain insights on the fundamental QoE relationships in the data and causes a lack of trust in such models, which inhibits the large-scale use of data-driven QoE models by researchers and providers.

With the recent advances in the field of eXplainable Artificial Intelligence (XAI), it is now possible to realize interpretable ML-based QoE models, and thus, increase the trust between stakeholders and QoE model. These advances include a large range of XAI techniques, which can be applied on top of existing black-box models, but also novel, sophisticated ML models, which are interpretable by design. The usage of XAI for QoE modelling is beneficial for several reasons. Besides speeding up the entire modelling process, data-driven QoE modelling allows to identify the most relevant QoE factors and their fundamental relationships to the Mean Opinion Score (MOS). At the same time, it helps to avoid introducing preferences or biases from different research teams and datasets into the model. All that is required is a large dataset with features and labels, i.e., descriptions of experiment conditions and stimuli (features) and the corresponding QoE ratings (labels). Datasets from different studies can also be merged to obtain better generalizable QoE models, and models can be refined automatically over time when new QoE studies have been conducted and new data become available.

In this work, we give an introduction to XAI and show the feasibility of data-driven QoE modelling with XAI on the use case of video streaming QoE. We choose the video domain since it has been well-researched in the QoE community in the past, e.g., in [3] and [4], and since this domain knowledge helps us to validate our insights. We evaluate different ML and XAI models and techniques on publicly available QoE data with respect to performance and explainability. We also compare existing expert video QoE models to these data-driven models. Moreover, we show that it is possible to identify QoE influence factors and model their impact on the MOS using XAI. Even though we evaluate the feasibility of data-driven QoE modelling on video streaming QoE in this work, transferring our approach to other domains like gaming or speech is straightforward. Finally, we discuss opportunities and challenges of data-driven QoE modelling.

This work is structured as follows: related work and XAI are discussed in Section II and Section III. Our video streaming QoE case study is presented in Section IV. Section V discusses opportunities and challenges, before Section VI concludes.

II. Related Work

Several QoE models have been proposed for traditional Internet services like video streaming [3], [4]. As shown in [3], there are various video QoE models, which consider different inputs, e.g., media-layer inputs like decoded audio/video signal, parametric inputs like packet headers, bitstream information like quantization parameter, as well as hybrid models, which consider a combination of the aforementioned inputs. In this work, we compare the performance of our data-driven QoE models to the expert QoE models P.1203 [2], Hoßfeld [5], Liu [6], Mao [7], Mok [8], Petrangeli [9], and VsQM [10]. The majority of these models considers the total stalling length, the number of stalling events, and the visual quality (bitrate and/or resolution) as important QoE influence factors. P.1203, Mao, Mok, and VsQM also factor in the initial delay. P.1203 additionally considers the number of video quality switches and it is the only expert model, which also incorporates ML. In [11], the authors showed that many of those QoE models perform significantly different as they attach different weights to the QoE influence factors. This suggests that a data-driven approach to QoE modelling can bring benefits to the QoE community in terms of model accuracy and generalizability.

The survey in [12] shows that ML-based QoE modelling in multimedia systems is already widely used, including Virtual Reality, 360 degree video, and gaming. However, the QoE models are based on shallow learning methods, e.g., Support Vector Machines (SVM), or on deep learning methods, which lack explainability. Thus, it is difficult to understand what QoE factors are relevant and how they affect the QoE score. To the best of our knowledge, explainable data-driven QoE modelling has not been addressed in QoE research yet. Thus, we are the first to evaluate its feasibility on a realistic use case and to discuss the deployment of XAI-based QoE modelling.

III. XAI: eXplainable AI

A general overview on XAI is provided in [13] and an extensive survey on XAI methods as well as a taxonomy for XAI methods in general can be found in [14].

XAI methods can be classified into techniques which explain a model locally, i.e., providing explanations for a single stimulus in terms of QoE factors and QoE rating, or globally, i.e., providing general reasoning for how a model derives the QoE rating from QoE factors. Additionally, XAI methods can be classified into post-hoc explainers and interpretable models.

Post-hoc explainers [14] are usually utilized to explain various black-box models, e.g., neural networks or ensemble techniques, after they have been trained. A widely used post-hoc explainer is SHAP values [15]. SHAP values originate from game theory and quantifies the contribution of each feature to the prediction by considering all potential feature subsets and learning a model for each feature subset. Other post-hoc explainers are LIME and Anchors, which have the drawback that they are usable for classification tasks only.

Interpretable models provide an explanation for how the model obtained the output by design. Prevalent models are, for example, the well-known linear models and decision trees, as well as the less known generalized additive models (GAM).

A GAM is a generalized linear model in which the model output is computed by summing up each of the arbitrarily transformed input features along with a bias [16]. Thus, GAMs can be described by the equation $g(\mathbb{E}[y]) = \beta + \sum_{i=1}^{n} f_i(x_i)$, where x_i is the i^{th} input feature of n total features, f_i is a univariate arbitrary predictor function for feature i, y is the target variable, and g is the link function. The link function g relates the expected value of the target variable to the learned predictor functions f_1 to f_n. The form of a GAM enables a direct interpretation of the model by analyzing the learned functions f_1 to f_n and the transformed inputs, which allows to estimate the influence of a feature. In this work, we utilize two state-of-the-art ML-based GAM models to model video streaming QoE, namely, Explainable Boosting Machine (EBM) [17] and Neural Additive Model (NAM) [16]. While EBM uses decision trees to learn the functions f_1 to f_n and gradient boosting to improve training, NAM utilizes arbitrary neural networks to learn the functions f_1 to f_n, resulting in a neural network architecture with n sub-networks. EBM extends GAM by also considering additional pairwise feature interaction terms, which are added on the right side of the GAM equation, while maintaining explainability.

IV. Use Case

Next, we evaluate the feasibility of explainable data-driven QoE modelling for the video streaming QoE use case.

Dataset: We crawled various publicly available sources for video QoE databases meeting specific criteria. In particular, we looked for databases containing rich information about experiment conditions and stimuli, which could have an impact on video streaming QoE, and thus, can be used as features for the ML models. This includes, for example, various video streaming KPIs, such as stalling information and visual quality. Moreover, the database was required to provide the corresponding subjective user ratings for all stimuli, which we require as labels to train ML models in a supervised fashion.

Five databases [2], [18]–[21] matched our criteria, and we preprocessed and aggregated them into a new dataset consisting of 2571 video streaming sessions. We characterize the databases shortly in Table I and observe that our new, merged dataset is highly heterogeneous due to the different databases. This can be seen, for example, in the duration of the video clips ranging from 10 up to 240 seconds. Finally, the MOS for the Waterloo databases [18], [19] is significantly higher compared to the other three databases.

For each video of a database, we extracted the initial delay in seconds, the number of stalling events, the total stalling duration in seconds (without initial delay), the playback duration in seconds (excluding stalling and initial delay), and the video width and height in pixels, and used them as features to describe the stimulus. Additionally, we extracted the bitrate (Mbps), the frame rate, and the stalling duration for every second of the video duration. Considering the distributions of these per-second KPIs in a single session, we computed the

TABLE I: Characteristics of considered video QoE databases.

Database	# Sessions	# Videos	Duration [s]	Initial Delay Incl.	Stalling Incl.	Avg. Stalling Length [s]	# Avg. Stallings	Avg. Bitrate [Mbps]	MOS (Std.)
Waterloo III [18]	450	20	10	100%	49.1%	3.2	2.4	2.52	3.43 (0.62)
Waterloo IV [19]	1350	5	28	12.3%	23.9%	3.6	2.0	4.60	3.65 (0.64)
LIVE Netflix I [20]	112	14	60-87	62.5%	50.0%	8.4	1.3	0.24	3.00 (0.70)
LIVE Netflix II [21]	420	15	25	0%	39.5%	2.7	2.5	0.60	2.99 (0.82)
ITUT-T Rec. P1203 [2]	239	157	60-240	30.1%	43.9%	12.9	1.4	1.77	3.15 (0.97)
Merged dataset	2571	211	10-240	29.48%	33.87%	4.8	2.1	3.13	3.43 (0.76)

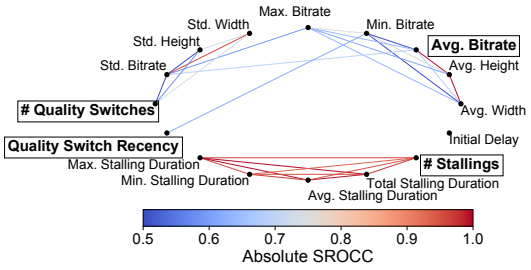

Fig. 1: Graph-based feature selection using cliques. Framed features are selected for data-driven feature set.

mean, standard deviation, minimum, and maximum statistics, which we also use as features. Finally, we also extract the number of quality switches, and the recency time of the last quality switch, but not the quality levels themselves since they are not compatible across the databases due to different minimum and maximum resolutions. Constant features, e.g., the frame rate statistics, were immediately removed from the feature set. Finally, we consider the continuous MOS of each stimulus as label, which shall be predicted by the QoE models. Note that we normalize all MOS scores across all databases.

Feature Selection: Selecting the most relevant features from a list of features is easy when you already have domain knowledge about the problem at hand. However, for unknown domains, e.g., for future multimedia applications, it is unclear which features are relevant QoE factors. Thus, we present a data-driven feature selection method based on graph theory, which selects uncorrelated features that are desired for XAI. We will compare the resulting feature set to an expert feature set, which is created based on domain knowledge.

1) Expert: For the expert feature set, we use video QoE domain knowledge derived from the QoE models presented in Section II. In particular, we look at the input parameters of these expert models and select the five most common parameters as features. These expert features include the average bitrate, the initial delay, the number of stalling events, the total stalling duration, and the number of quality switches.

2) Data-Driven: To select uncorrelated features in a data-driven fashion, we utilize a custom, simple, non-optimized graph-based method similar to [22]. First, we compute Spearman's rank correlation coefficient (SROCC) for all features. Using the SROCC, we build a graph by interpreting each feature as a node. Two nodes are connected in the graph if their absolute correlation is higher than a threshold x. In this work, we add an edge between two nodes only if $|SROCC| > 0.5$. Figure 1 depicts the resulting graph for our dataset. Edges between features are colored according to the absolute intensity of the SROCC. We can easily see that all stalling-related features are clustered together and do not share an edge with other features. To find the highly correlated subsets, we then compute all maximal cliques, i.e., fully-meshed subsets of the graph, and sort the cliques by size in a descending fashion. We then iterate through each clique and select the feature with the highest SROCC to the MOS, add the feature to our feature set, and blacklist all other features in the clique. After all cliques have been checked, we obtain the final feature set. For our exemplary dataset, four features are selected as most indicative. These are the average bitrate, the number of stalling events, the number of quality switches, and the recency time of the last quality switch. When comparing the expert and the data-driven feature set, we observe that both feature sets contain the average bitrate, the number of stallings, and the number of quality switches. Thus, there is a large agreement in both sets, while our data-driven method excludes correlated features, which validates the approach.

Model Comparison: We perform a comparison between selected expert video streaming QoE models and ML models, as well as between our expert and data-driven feature sets. As baseline black-box ML models, we use XGBoost (XGB) [23], a tree-based boosting ensemble method, which has proven to be particularly suited for tabular data, as well as Random Forest (RF) and Deep Neural Networks (DNN). Nevertheless, we will explain these black-box models using the post-hoc method SHAP. In addition, we utilize three different inherently interpretable XAI models in this work, namely, Decision Tree (DT), Explainable Boosting Machine (EBM), and Neural Additive Model (NAM). For the training of all regression models, we perform extensive hyperparameter tuning. We explicitly do not split our dataset into train and test set since our dataset is too small and generalizability is not our goal here. Instead, we aim to train models providing high performance and high quality explanations for their internal decision making.

Table II shows the performance of each expert and ML model per feature set with respect to common metrics, which can describe the goodness of fit for regression models. These are root mean square error (RMSE) and mean absolute error (MAE) on the MOS range, which should be as small as possible, and the coefficient of determination (R^2), which should be close to 1. The table shows that almost all ML-based models outperform even the best expert model P.1203 in our dataset. However, this is no surprise since our ML models are fitted directly to our dataset. Nonetheless, the expert QoE models could have been expected to generalize

TABLE II: Performance comparison of expert and data-driven QoE models for expert and data-driven feature sets.

Features	Data-Driven			Expert		
Metric	RMSE	MAE	R^2	RMSE	MAE	R^2
P.1203	-	-	-	0.61	0.47	0.35
Hoßfeld	-	-	-	1.62	1.44	-3.57
Liu	-	-	-	1.77	1.63	-4.45
Mao	-	-	-	0.73	0.61	0.35
Mok	-	-	-	0.88	0.73	-0.35
Petrangeli	-	-	-	1.62	1.44	-3.57
VSQM DASH	-	-	-	1.60	1.41	-3.49
XGB	0.25	0.15	0.89	0.23	0.15	0.90
RF	0.28	0.21	0.86	0.27	0.19	0.87
DNN	0.47	0.37	0.61	0.46	0.35	0.63
DT	0.61	0.47	0.35	0.63	0.49	0.31
EBM	0.43	0.34	0.67	0.42	0.32	0.69
NAM	0.51	0.40	0.54	0.50	0.39	0.56

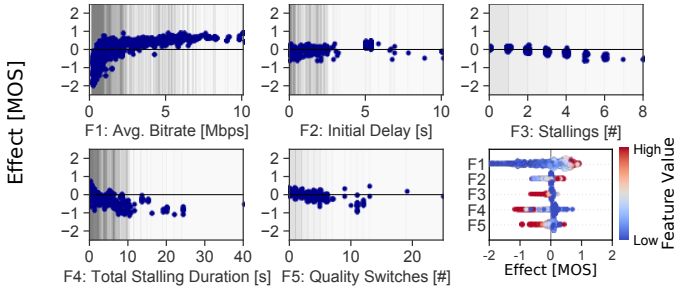

Fig. 2: Effect of expert features on MOS with SHAP values.

well over unseen data. For both feature sets, we can also see only marginal differences in both directions. This indicates that the data-driven choice of the feature set was an excellent one for our dataset. Finally, we observe that the explainable models DT, EBM, and NAM all perform worse than the black-box models XGB and RF, indicating that explainable models suffer a minor performance loss.

Explainability: Next, we want to explain the models' decisions. Here, we use only the expert feature set as it contains an additional feature, and thus, allows for more insights.

1) SHAP Values: While the models DT, EBM, and NAM are explainable by design, our black-box models can only be explained using post-hoc explainers. In the following, we explain the internals of XGB with SHAP values. Figure 2 shows how MOS or SHAP values, respectively, are affected by different feature values. The gray shades in the background denote the density of the data, i.e., the normalized number of samples available for a specific feature value. Darker shades correspond to higher densities. The obtained SHAP value on the y-axis represents how a single instance with a given feature value differs from the average of the entire dataset. For example, stimuli with an average bitrate below 1 Mbps show a MOS that is around 1-2 points lower on average than the average MOS of all stimuli. Although SHAP plots resemble traditional main effect plots, they have the advantage of implicitly accounting for feature interactions.

We notice that the SHAP values of the average bitrate show a strong monotonic trend, thus indicating the importance

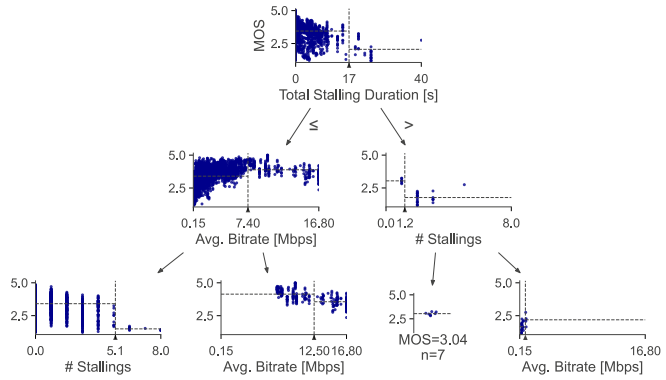

Fig. 3: Top layers of best decision tree for prediction of MOS.

of this QoE factor. When we consider the initial delay, we observe SHAP values centered around 0, i.e., negligible impacts on MOS. For the number of stalling events and the number of quality switches, there are also only minor negative linear trends visible. In contrast, the total stalling length has a stronger negative impact with increasing stalling durations, which comes as expected. On the bottom right, a summary over the SHAP values per feature is shown, which confirms the previously observed trends. We remark that the SHAP values in our figures are highly variant, and thus, not very accurate and difficult to interpret in mathematical terms. Note also that SHAP values require an additional computing step after the training and that the computational complexity of SHAP values increases dramatically with the number of features and samples. Thus, we argue that inherently explainable models may be the better choice for data-driven QoE modelling.

2) Decision Tree (DT): Decision trees are easy to interpret, because the learned decision rules can be visualized in an if-else fashion. Figure 3 visualizes the decision rules for a pruned decision tree using the Python library *dtreeviz*. On each tree level, the figure shows at which feature value the model performed the split (triangle), thereby directly explaining the internals of the tree. For example, it can be seen that the model first splits the dataset based on a total stalling duration threshold of 17s. Afterwards, the model uses the average bitrate or the number of stallings to divide the dataset further. This continues down the tree until a final prediction is made at the leaves. We can, for example, see that the model classifies a sample with a total stalling duration above 17s and no more than one stalling event with a MOS of 3.04. Thus, decision trees are easily explainable, however, the usually poor performance (cf. Table II) comes as a drawback here.

3) Generalized Additive Models (GAM): Now, we consider EBM and NAM. Figure 4 shows the learned predictor functions for the best set of hyperparameters for both EBM (green) and NAM (blue). The findings are again mostly in line with SHAP values, but both models provide much smoother shape functions and are thus easier to interpret. EBM and NAM differ only marginally and mostly in areas where the data density is low. Here, EBM outperforms NAM (cf. Table II) by overfitting on single data points using feature interaction

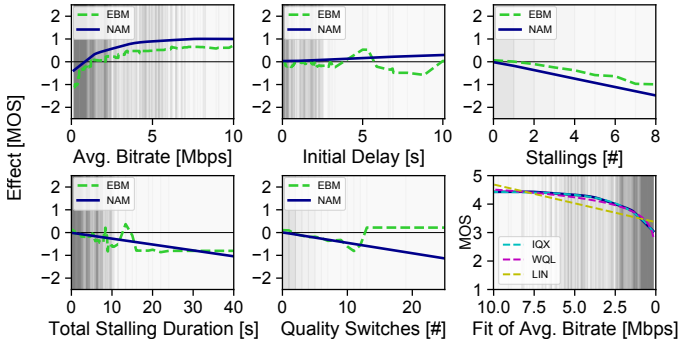

Fig. 4: Effects of expert features on MOS with Explainable Boosting Machine (EBM) and Neural Additive Model (NAM).

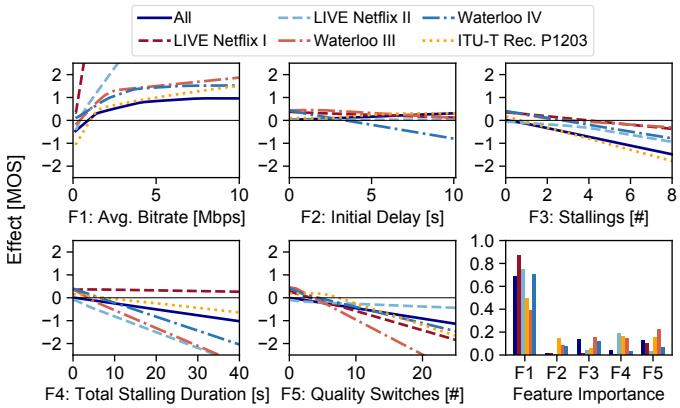

Fig. 5: Effects of expert features on MOS for the entire dataset and all databases composing the dataset with NAM.

terms. We can see this, for example, for a high total stalling duration and a high number of quality switches, where at some point EBM stops the negative trend and strongly contrasts its previous trend to improve predictions for extreme outliers. Using the smooth predictor functions, it is easy to apply curve fitting. In the bottom right plot, we fit the average bitrate predictor function of NAM, which was shifted by the average MOS of the dataset to obtain the original MOS scale on the y-axis, on an inverted x-axis using exponential (IQX), logarithmic (WQL), and linear functions (LIN). Note that this constitutes a univariate mapping of average bitrate to MOS, neglecting the other influencing factors. We observe that our predictor function follows the WQL hypothesis [5] (red) with a high $R^2 = 0.967$. This is in line with the mechanics of P.1203, where the authors of [24] showed the same logarithmic behavior for the bitrate in mode 0. Summarizing, using GAMs, we obtain valuable easy to interpret functions, which explain fundamental relationships between QoE factors and MOS.

Impact of Databases: Finally, we analyze the impact of the different databases on the outcomes of what a model, here NAM, learns. Fig. 5 depicts again the predictor functions learnt by NAM for each feature and for each database. First of all, we observe that the predictor functions can diverge strongly for some features. This can be seen for example for the average bitrate, where the predictor functions for the LIVE Netflix I and LIVE Netflix II databases strongly differ from the other databases. We attribute this divergence to the characteristics of the considered databases (cf. Table I). Next, a similar linear trend is visible for most of the other features and databases. Here, the linear trends differ only in their slope. Further, for the LIVE Netflix I database, the total stalling duration does not show a negative trend, but a small positive influence, thereby contradicting the other databases. This heterogeneity of the input also explains why some of the expert QoE models struggled with performing well on our merged dataset. Finally, we look at the normalized feature importance scores, i.e., absolute feature output relative to sum of all absolute feature outputs, in the bottom right of the figure. All databases agree on the average bitrate as most important feature, while differences between the databases are visible for the other features.

V. Discussion

Apart from speeding up modelling and avoiding to introduce modelling biases, the biggest advantage of data-driven QoE modelling is its higher accuracy and generalizability compared to manual QoE models. We could observe this also in our video streaming QoE case study. These advantages stem from the fact that ML-based models are not limited to certain classes of continuous, well-behaving functions, which are typically used in manual modelling. However, the challenge with ML-based models is to avoid overfitting, where the model is sensitive to noise, but misses the underlying relationships in the data. Overfitting can typically be avoided by model regularization or collecting sufficiently large datasets.

To successfully apply data-driven QoE modelling, a purposeful data collection is key. It has to ensure that all (or at least the most important) QoE factors are included in the dataset on their full parameter range with a sufficient number of samples. While it is easy for controlled lab/crowdsourcing studies to define the feature values, constraints on the study budget (time, cost) limit data collection to a small set of selected feature values. In contrast, field studies can cover all feature values that appear in the wild, however, they will only collect few data samples for rare events, e.g., video sessions with many stalling events. To avoid data bias, feature values should be balanced, which might require to purposefully generate rare events in the field. Additionally, we require a thorough data cleaning. While it is possible to impute missing features due to measurement errors, they increase the risk of inserting a bias. Thus, filtering out missing or strange feature values should be preferred.

In our case study above, ML-based models could use a larger dataset compared to the expert QoE models. However, adding new data and retraining an ML model is natural and easily possible for data-driven modelling, which is an advantage in the long run. Eventually, data-driven QoE models would be able to cope with concept drift, i.e., changes in the importance of influence factors over time, e.g., due to changed expectations of end users. The challenge here is that QoE studies are rarely conducted temporal and population-based snapshots, such that we cannot frequently update the model.

Ideally, a pipeline could be implemented, which provides a continuous stream of features and QoE ratings to enable online learning and keep the QoE models up to date. While this is difficult for research endeavours, service providers could include such QoE feedback streams into their applications.

Comparing black-box and interpretable ML models, there is a slight trade-off between performance and explainability. However, as shown above, it should be negligible in the context of QoE modelling. Instead, XAI allows to fully understand the model decisions, identifying relevant QoE factors and their relationships to the QoE score. Nevertheless, it has to be considered that explaining models becomes inherently more difficult when the number of input features increases. Highly correlated features and interactions may further lead to misinterpretations when using XAI since the influence of a feature may also depend on other features. To obtain reliable and trustworthy explainable models, it is therefore crucial to exclude highly correlated features.

Finally, although we demonstrated XAI-based QoE modelling only for video streaming from a research perspective, it is important to understand that the whole process is easily applicable in other domains like speech or gaming. Apart from that, it can also be highly beneficial for providers of services and networks to use XAI when implementing a continuous QoE monitoring. They could integrate visualizations of trends like Figure 4 into dashboards, thus, allowing to easily obtain a deeper understanding of the QoE in their system.

VI. Conclusion

In this work, we propose data-driven QoE modelling using XAI. It allows to obtain valuable insights directly from the data without cumbersome and time-consuming manual modeling. XAI-based QoE models are expected to avoid modelling biases and generalize better to new or unseen data.

We demonstrated the feasibility of explainable data-driven QoE model by conducting a case study for video streaming QoE. Without any manual modelling, we were able to select the most important features, identify the most relevant QoE factors, and could describe the fundamental relationships, which govern their influence on the QoE scores. Our findings agree with previous results from the QoE community, validating our approach. Moreover, we compared a variety of black-box and interpretable ML models with expert video streaming QoE models. ML-based models were able to better generalize over all data in our heterogeneous dataset, while expert models struggled with performing well, as they were designed for a subset of the data only. Finally, we discussed opportunities and challenges related to data-driven QoE modelling.

To sum up, we find that, due to technical advances, data-driven explainable QoE modelling is ready for deployment. Thus, researchers and providers of services and networks should be interested in deploying XAI-based QoE modelling to obtain a better and more general understanding of QoE impact factors and their relationships to end users' subjective experience. This will allow to improve services and networks in terms of QoE, thus, avoiding user churn and revenue losses.

Acknowledgement

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant SE 3163/3-1, project number: 500105691. The authors alone are responsible for the content.

References

[1] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi *et al.*, "Qualinet White Paper on Definitions of Quality of Experience," 2013.

[2] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia *et al.*, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P. 1203: Open Databases and Software," in *ACM MMSys*, 2018.

[3] N. Barman and M. G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges," *IEEE Access*, 2019.

[4] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE COMST*, 2015.

[5] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience," in *Data Traffic Monitoring and Analysis*, 2013.

[6] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and Validating User Experience Model for DASH Video Streaming," *IEEE Trans. Broadcast.*, 2015.

[7] H. Mao, R. Netravali, and M. Alizadeh, "Neural Adaptive Video Streaming with Pensieve," in *ACM Special Interest Group on Data Communication*, 2017.

[8] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the Quality of Experience of HTTP Video Streaming," in *IFIP/IEEE IM*, 2011.

[9] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. De Turck, "QoE-Driven Rate Adaptation Heuristic for Fair Adaptive Video Streaming," *ACM TOMM*, 2015.

[10] D. Z. Rodríguez, R. L. Rosa, E. C. Alfaia, J. I. Abrahão, and G. Bressan, "Video Quality Metric for Streaming Service using DASH Standard," *IEEE Trans. Broadcast*, 2016.

[11] A. Seufert, F. Wamser, D. Yarish, H. Macdonald, and T. Hoßfeld, "QoE Models in the Wild: Comparing Video QoE Models Using a Crowdsourced Data Set," in *IEEE QoMEX*, 2021.

[12] 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*, 2022.

[13] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[14] A. B. Arrieta, N. Díaz-Rodríguez *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information fusion*, 2020.

[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NIPS*, 2017.

[16] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural Additive Models: Interpretable Machine Learning with Neural Nets," *NIPS*, 2021.

[17] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A Unified Framework for Machine Learning Interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[18] Z. Duanmu, A. Rehman, and Z. Wang, "A Quality-of-Experience Database for Adaptive Video Streaming," *IEEE Trans. Broadcast*, 2018.

[19] Z. Duanmu, W. Liu, Z. Li, D. Chen, Z. Wang, Y. Wang, and W. Gao, "Assessing the Quality-of-experience of Adaptive Bitrate Video Streaming," *arXiv preprint arXiv:2008.08804*, 2020.

[20] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of Temporal Effects on Subjective Video Quality of Experience," *IEEE TIP*, 2017.

[21] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards Perceptually Optimized End-to-End Adaptive Video Streaming," *arXiv preprint arXiv:1808.03898*, 2018.

[22] D. T. Schroeder, K. Styp-Rekowski, F. Schmidt, A. Acker, and O. Kao, "Graph-Based Feature Selection Filter Utilizing Maximal Cliques," in *IEEE SNAMS*, 2019.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *ACM SIGKDD*, 2016.

[24] M. Seufert, N. Wehner, and P. Casas, "Studying the Impact of HAS QoE Factors on the Standardized Qoe Model P. 1203," in *ICDCS*, 2018.