IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 7, 2019, accepted May 29, 2019, date of publication June 3, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920477

No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications

NABAJEET BARMAN[1], (Member, IEEE), EMMANUEL JAMMEH[1,2], (Member, IEEE), SEYED ALI GHORASHI[1,3], (Senior Member, IEEE), AND MARIA G. MARTINI[1], (Senior Member, IEEE)

[1]Wireless Multimedia & Networking Research Group (WMN), Kingston University London, Kingston upon Thames KT1 2EE, U.K.
[2]School of Computing, Electronics, and Mathematics, University of Plymouth, Plymouth PL4 8AA, U.K.
[3]Cognitive Telecommunication Research Group, Department of Telecommunications, Faculty of Electrical Engineering, Shahid Beheshti University, G.C., Tehran 1983969411, Iran

Corresponding author: Maria Martini (m.martini@kingston.ac.uk)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie Grant under Agreement 643072 and in part by the Kingston University's ISC Fund.

ABSTRACT Recent years have seen increasing growth and popularity of gaming services, both interactive and passive. While interactive gaming video streaming applications have received much attention, passive gaming video streaming, in-spite of its huge success and growth in recent years, has seen much less interest from the research community. For the continued growth of such services in the future, it is imperative that the end user gaming quality of experience (QoE) is estimated so that it can be controlled and maximized to ensure user acceptance. Previous quality assessment studies have shown not so satisfactory performance of existing No-reference (NR) video quality assessment (VQA) metrics. Also, due to the inherent nature and different requirements of gaming video streaming applications, as well as the fact that gaming videos are perceived differently from non-gaming content (as they are usually computer generated and contain artificial/synthetic content), there is a need for application-specific light-weight, no-reference gaming video quality prediction models. In this paper, we present two NR machine learning-based quality estimation models for gaming video streaming, NR-GVSQI, and NR-GVSQE, using NR features, such as bitrate, resolution, and temporal information. We evaluate their performance on different gaming video datasets and show that the proposed models outperform the current state-of-the-art no-reference metrics, while also reaching a prediction accuracy comparable to the best known full reference metric.

INDEX TERMS Quality assessment, no reference, gaming video streaming, machine learning, regression, quality of experience, video quality metrics.

I. INTRODUCTION

Gaming video streaming has gained much popularity in recent years, due to the advances made in the field of both passive and interactive services. Interactive gaming streaming applications or cloud gaming, as popularly known, refer to applications where the user's gameplay is processed in the cloud. The user receives the gameplay which is then rendered on a screen based on which users can input gameplay commands. The passive scenario, on the other hand, refers to Over-The-Top (OTT) services, such as Twitch.tv and YouTubeGaming, where a viewer can watch videos of the gameplay of other players. Such passive OTT gaming video streaming services have seen tremendous growth, in terms of both number of viewers and the number of streamers. For example, Twitch.tv alone currently has over 15 million streamers and over nine million daily active users and is ranked 4th in terms of peak Internet traffic in the US, just behind Netflix, YouTube, and Apple [1].

With the ever increasing demand for multimedia services, as well as increasing user expectations of content availability anywhere, anytime, anyplace, there has been a recent shift from traditional Quality of Service (QoS) based management towards Quality of Experience (QoE) based management

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Ji.

of multimedia services. For the continued success of such services, it is imperative that the end users' perceived quality is accurately estimated so that it can be managed optimally so as to ensure the best possible gaming video quality delivery.

This is usually performed via subjective tests [2]. The disadvantage of subjective testing is that it is a time consuming and expensive process and is not suitable for many applications such as real-time network monitoring/resource allocation. To overcome the shortcomings of subjective tests, there has been a growing interest in objective quality metrics/models which predict the quality of images and videos as perceived by the end users (e.g., Peak Signal to Noise Ratio (PSNR) [3], Structural Similarity (SSIM) [4], Video Multimethod Assessment Fusion (VMAF) [5]). The performance of such quality metrics is evaluated based on various measures, but most importantly based on the correlation between their estimation of quality with subjective scores obtained from subjective quality assessment. Despite their poorer performance, objective metrics are preferred due to their speed and practicality. Image Quality Assessment (IQA) as well as Video Quality Assessment (VQA) metrics are classified as Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) depending on the amount of reference information used in quality estimation [6]. FR metrics compare complete reference information with information from the distorted signal to provide an estimate of the quality of the received signal. RR metrics use a part of the reference information while No-Reference (NR) metrics do not use any reference information.

Over the past two decades, researchers have investigated methods and techniques to estimate audio, image and video quality as perceived by the end users. However, gaming videos are generally different from non-gaming content because they are usually computer generated, contain artificial/synthetic content and are perceived differently by users [7]. In fact, the studies in [8] and [7] reported differences between gaming and non-gaming video quality for the same encoding process. In [7] the authors found that some of the most popular and widely used quality assessment metrics resulted in a lower correlation between predicted quality and actual quality for gaming videos as compared to that for natural videos. Most of the IQA and VQA metrics, such as Blind Image Quality Index (BIQI) [9], Natural Image Quality Evaluator (NIQE) [10] and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [11]) are designed taking into account statistics or properties which are inherent to natural images. To investigate the performance of such metrics for gaming video quality assessment, the authors evaluated in [12] the performance of eight widely used and popular quality assessment metrics (3 FR, 2 RR, and 3 NR) using the GamingVideoSET dataset proposed in [13]. They found that VMAF [5] performs best in terms of both Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Correlation Coefficient (SROCC). Although NIQE performs best among the NR metrics such as BIQI and BRISQUE, its performance is still unsatisfactory for practical applications and well below the state-of-the-art FR metrics.

While FR quality metrics, in general, perform better than RR and NR metrics, there exist some applications where there is no reference signal available. Gaming video streaming is one of such applications since, due to the inherent nature of the service, reference information is not available. Therefore, for such applications, FR and RR metrics cannot be used to predict/estimate the quality. Hence, the availability of custom NR metrics for gaming content with high performance is necessary for continued success and further improvement of existing services.

In the absence of any NR gaming video quality metric/model that meets existing requirements of high accuracy and low computational complexity to estimate accurately the quality of gaming videos in real time, in this paper we present two machine learning based lightweight[1] gaming video quality estimation models. Both models, due to their low complexity nature, can therefore be used as the first stage of an optimized online gaming QoE management system, even on thin clients. The main contributions of the paper are as below:

1) We propose a Neural Network (NN) based No-Reference Gaming Video Streaming Quality Index (NR-GVSQI). The model is designed using subjective ratings (Mean Opinion Score (MOS)) from two open-source datasets. The proposed model is shown to outperform existing state-of-the-art NR metrics.
2) We also present a Support Vector Regression (SVR) based model, No-Reference Gaming Video Streaming Quality Estimator (NR-GVSQE), which is designed using FR VQA scores from GamingVideoSET. Our test of the proposed model, NR-GVSQE, on an unseen dataset shows that the proposed model, although no-reference, results in almost the same performance as the state-of-the-art full-reference VQA metric, VMAF.
3) Additionally, this paper presents an open source dataset, KUGVD, which consists of both subjective (MOS) ratings and objective analysis considering six gaming videos.

The rest of this paper is organized as follows. Section II describes the previously proposed NR machine learning (ML) based QoE models. In Section III we briefly describe the existing open source dataset, GamingVideoSET, and also introduce our newly designed dataset, KUGVD. Section IV describes the extracted features and the feature selection methods along with the model development methodology. Section V describes the development, testing and validation of the NR-GVSQI model to predict the MOS scores obtained via subjective tests (MOS). Section VI describes the NR-GVSQE model which is developed using an existing state-of-the-art FR VQA metric (VMAF) as the target output.

[1] By lightweight, we refer to the fact that all the features used in this work can be extracted in real-time without the need for high computational power and hence can be used for real-time quality monitoring.

FIGURE 1. Some of the sample videos used in this work. (a) Counter Strike: Global Offensive. (b) FIFA 2017. (c) H1Z1: Just Kill. (d) League of Legends. (e) Hearthstone. (f) Overwatch.

Section VIII concludes the paper with a summary of key findings and possible future work.

II. RELATED WORK

With the advancements in the field of machine learning, the field of quality assessment in recent years has seen many proposed quality metrics/models based on machine learning algorithms, using different types of quality impacting factors, such as jitter, packet loss, compression artifacts (blockiness, blurriness, flickering, etc.) and rescaling. Since this study is focused solely on the design of NR metrics, we provide a brief review of recent works which have used ML algorithms to predict image/video quality without using any reference information.

In one of the earliest works in this direction, the authors in [14] used a Back-Propagation Artificial Neural Network (BP-ANN) to estimate the PSNR of H.264/AVC encoded video and obtained 97.8% correlation between the predicted and the actual PSNR. However, PSNR has been shown not to correlate well with QoE [3], [15]. Jiang *et al.* in [16] used a three-layer BP-ANN to predict the quality of high definition video, using features such as image blur, entropy, blocking artifacts, frequency energy, chroma information, and temporal information. Choe *et al.* used a three-layer BP-ANN to predict subjective quality scores based on features that were extracted from the H.264 bit-stream on a frame-by-frame basis. The proposed method used features based only on compression impairments and not on network QoS. The authors in [17] used an adaptive network-based fuzzy inference system based hybrid ANN to train a NN to estimate the quality of video transmitted over a wireless local area network and universal mobile telecommunication system. The prediction model used content type, frame rate and sender bitrate as application layer parameters and block error rate and link bandwidth as physical layer parameters. Shahid *et al.* in [18] used a 2-layer BP-ANN to predict the PSNR, Perceptual Evaluation of Video Quality (PEVQ) and SSIM based on features such as bits per frame, percentage of inter blocks, average motion vector length, and average Quantization Parameter (QP). Although, PSNR and PEVQ were accurately predicted, the SSIM score is predicted with less accuracy. Wang *et al.* in [19] used features such as picture size, bitrate (BR), frame rate, Group of Pictures (GOP) structure, picture type, macroblock type, QP, motion vectors, coded block pattern, and Discrete Cosine Transform (DCT) coefficient as inputs to a 3-layer BP-ANN for quality assessment of MPEG-2 video streams. However, they did not compare their method with other regression methods and they did not consider feature selection or Principal Component Analysis (PCA). Cherif *et al.* in [20] used features such as QP, base-layer loss rate, enhancement layer 1 and layer 2 loss rate as inputs to a 3-layer BP-ANN to estimate the QoE of H264/SVC bit stream.

Khattabi *et al.* [21] used a BP-ANN with 3 hidden layers, and features such as the average of differences, the standard deviation of discrete Fourier transform differences, the average and standard deviation of Discrete Cosine Transform (DCT) differences, the variance of color energy, luminance, and chrominance, to predict both MOS and PSNR. The complexity was high, due to the high number of features as well as the NN structure, and the authors did not reduce the dimensionality. Singh *et al.* [22] used a 3-layer ANN for NR QoE monitoring of H.264/AVC encoded videos streamed using HTTP/TCP in the context of IPTV. In [23], the authors used SVR for quality prediction, compared the performance with different visual quality predictors and reported improvement in prediction accuracy. Sunala and Anurenjan [24] used bitrate, SSIM, and interframe transformation fidelity as

inputs to an ANN for video quality estimation. The authors in [25] used a radial basis function network for QoE estimation of video streamed over wireless networks, using cross-layer features such as bitrate, frame rate and resolution (RES) at the application layer, Packet Loss Ratio (PLR) at network layer, video content features and the screen size of the terminal equipment. Xue *et al.* in [26] introduced a NR ANN-based video quality metric to predict the quality by considering the impact of frame freezing due to packet loss and/or late arrival. They used features based on freezing events such as the number of freezes, freeze duration statistics, inter-freeze distance statistics, frame difference before and after the freeze, normal frame difference, and the ratio of them.

In [27], the authors used a low complexity Multilayer Perceptron (MLP) NN for video quality assessment for mobile streaming services which can be used in smartphones in 4G-LTE. In [28], delay, jitter, PLR, and mean loss burst size are used as the inputs to a three-layer BP-ANN to assess the QoE of video services in LTE networks. In [29], 16 features including blackout, blockiness, block loss, blur, brightness, contrast, exposure, flickering, freezing, interlacing, letterboxing, noise, pillar-boxing, slicing, spatial activity, and temporal activity are used as the inputs of a BP-ANN for high definition video quality assessment. In [30], PLR, the percentage of damaged frames and the percentage of different temporal classification frame which loses the packet, are used to train a feed-forward BP-ANN wireless video quality assessment model. In [31], first, a 2D convolutional NN is used to learn the spatial quality features at the frame level. Then, at the sequence level, the motion information is extracted as a temporal quality feature. A multi-regression model is then used for video quality measurement. In [32] the authors use restricted Boltzmann machine as an unsupervised deep learning method for video quality assessment. BR, number of frames, scene complexity, video motion, blur mean, blockiness, and motion intensity are used as features. They achieved an average of 78 to 91 percent correlation with well-known FR degradation assessment model VQM. In terms of scalability, they reported that only nine samples from the original video content types were sufficient to accurately assess the remaining of 864 videos of the dataset. More recently, the authors in [33] presented a FR and NR IQA metric that has a superior performance with respect to the state-of-the-art NR and FR IQA metrics when its performance was evaluated using three publicly available databases. The authors in [34] proposed a NR deep neural network IQA metric (MEON) consisting of two sub-networks each catering for two sub-tasks (distortion identification and quality prediction) for quality assessment with dependent loss functions. Their model is shown to achieve superior performance over the existing NR IQA metrics including the one proposed in [33] considering four different publicly available datasets. Inspired by the MEON model, a deep neural network based NR VQA model called V-MEON is proposed by the authors in [35] which provides an estimation of both quality scores as well as codec type. A comparison with existing NR metrics is shown to achieve high performance on two publicly available datasets.

Although there are many recent works in the field of quality assessment - as described above - most of these studies are limited in one or more of the following: different context (IPTV, etc.), very high complexity ([21]), older/different codecs (SVC, MPEG-2, etc.), evaluation methodology (few videos/single datasets, no subjective ratings, etc.), design for image quality assessment, rather than video quality assessment. Furthermore, all of these studies are limited to non-gaming content, whereas, as discussed earlier, gaming content has different streaming requirements and is inherently different from non-gaming content. Our work, on the other hand, focuses solely on gaming video content and uses two different datasets with stimuli representing compression artifacts as currently used by various OTT service providers.

III. DATASETS

In this work we use two datasets. One of these is the open source gaming video dataset GamingVideoSET [13]. Here we provide a brief discussion of the dataset and refer the reader to [13] for more details. The dataset consists of a total of 24 reference videos of 30 seconds duration, encoded in 24 different resolution-bitrate pairs to obtain 576 distorted (compressed) video sequences. In addition, MOS ratings for 90 stimuli conditions (six videos, 15 multiple resolution-bitrate pairs) are provided. MOS values are calculated as the average of the ratings provided by individual test participants during a subjective test for a particular video sequence.

TABLE 1. Overview of the two datasets used in this work.

Parameter	GamingVideoSET	KUGVD
Number of reference videos	24	6
Number of distorted videos	576	144
Games for Subjective Assessment	**CSGO**, FIFA, *H1Z1*, HS, LoL, PC	**CSGO**, FIFA, *H1Z1*, HS, LoL, OW
Number of Test Subjects	25	17
Subjective Test Environment Condition	ITU-R BT.500	ITU-R BT.500
Subjective Test Methodology	ACR	ACR
Number of Stimulus for Subjective Quality Assessment	90	90

The game in bold indicates the same gaming video across the datasets.
The underlined games indicates different games across the two datasets.

Since only 90 subjective ratings are available in the GamingVideoSET dataset, which may lead to overfitting of the data when building the model using subjective ratings, we created another dataset, Kingston University Gaming Video Dataset (KUGVD). In order to not include any new type of impairment to the dataset other than what the model would be trained on, we used the same encoding settings as in GamingVideoSET. We selected six gaming videos and encoded them in the same 24 resolution bitrate pairs as was done with the GamingVideoSET resulting in 144 stimuli. For subjective assessment, we selected 90 stimuli with the same resolution-bitrate pairs as in GamingVideoSET. Table 1 summarizes the parameters of the two datasets.

Since the subjective test was carried out at different places using a different set of participants, we decided to keep one game, CSGO, across both datasets, which then acts as anchor conditions (see Section III-C). Four of the games selected were the same (FIFA17, H1Z1, HS, and LoL) but a different part (scenario) of the game was considered. Depending on the stage/scenario of the game, the game content complexity can vary a lot. Hence, considering the same game but a different scenario (scene) will allow us to investigate whether the model designed using one dataset (considering a particular scenario) is robust enough to predict the quality with reasonable accuracy when considering a different scenario from the game. Additionally, we selected the game Overwatch (OW), a first-person shooting genre game, as it is more popular on Twitch.tv and is of high complexity. This allows us to introduce a totally unknown game in either of the datasets: Project Car (PC), a car racing game being the other one, with complexity and characteristics which will not be present during the training phase. This allowed us to design a robust model which can lead to satisfactory performance even when evaluating the quality of an unknown game type.

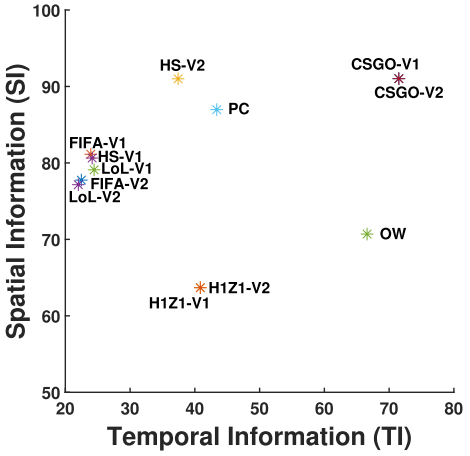

FIGURE 2. SI and TI plot for 12 gaming video sequences, six each from gaming video SET and KUGVD.

A. DESCRIPTION OF THE DATASETS

Spatial Information (SI) and Temporal Information (TI) as defined in ITU-T Rec. P.910 [36] are used as indicators of content complexity. Fig. 2 shows the SI vs. TI plots of the gaming videos considered for the subjective tests from both datasets. An interesting point to note is that the SI and TI for the video sequence from the game H1Z1 are the same even when the considered scenarios are different. LoL and FIFA are approximately of the same complexity while the HS sequence in KUGVD is of higher spatial and temporal complexity compared to the corresponding HS sequence in GamingVideoSET.

The videos were encoded at the same 24 resolution bitrate pairs (same as those used in GamingVideoSET, see Table 2) resulting in a total of 144 video sequences. Three resolutions and five bitrates from six videos from each dataset resulting in 90 stimuli were considered for subjective quality assessment which are shown in bold text in Table 2.

TABLE 2. Resolution-bitrate pairs of com- pressed video sequences.

Resolution	Bitrate (kbps)
1080p	**600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**
720p	**500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000**
480p	**300**, 400, **600**, 900, **1200**, **2000**, **4000**

B. TEST ENVIRONMENT AND SET UP

In line with the procedure followed by authors in [13], we conducted a subjective quality assessment test at Kingston University, London, United Kingdom in a test lab adhering to ITU-R Rec. BT.500 standard [2]. The display monitor used was a 55″ Samsung 4K monitor. The 480p and 720p videos were upscaled and then, together with 1080p videos, decoded to raw YUV format. These were then put into an *.mp4* container for playback at 1080p resolution at the center of the display monitor with the rest of the pixels of the display fully black. The playlist was randomized in order to avoid learning effects. For training, we selected 4 videos from two games which were not part of the test, so as to make the test participants familiar with the test interface and the rating tool. The test participants were tested for visual acuity and color blindness using Snellen charts and Ishihara plates, respectively. After removing the ratings from test subjects who failed either of the visual tests, a total of 17 valid subjective test ratings were obtained.

C. ALIGNING SUBJECTIVE TESTS SCORES

Since the subjective tests are conducted across different labs with different factors such as display, number and demographics of the test participants, the usual practice is to use anchor conditions (same test videos) across the different datasets and then use the MOS scores of these anchor conditions to determine a linear mapping function [37] which is then used to scale all MOS scores of the dataset(s). In our study, the gaming video sequence from the game CSGO is the same across both datasets (a total of 15 conditions taking into account three resolutions and 5 bitrates for each resolution). Considering the fact that GamingVideoSET contains MOS scores using more test participants as compared to KUGVD, we use the 15 MOS scores for CSGO from the GamingVideoSET as the reference scores for the anchor conditions and then use the linear mapping function $f(x) = mx + b$ as proposed in [37] to obtain the mapping between the anchor conditions. Using MOS scores of the anchor conditions, the coefficients m and b of the mapping function are obtained to be 0.9254 and -0.2613 respectively. Fig. 3 shows the scatter plot for the MOS scores and the linear fit. The goodness of fit scores obtained are as follows: SSE: 0.7114, R-square: 0.9443, Adjusted R-square: 0.94 and RMSE: 0.2339, indicating a good fit between the anchor MOS scores. The correlation between the anchor MOS conditions is obtained as 0.9875. As the fit is linear, there is no effect of the scaling of the MOS scores of KUGVD on

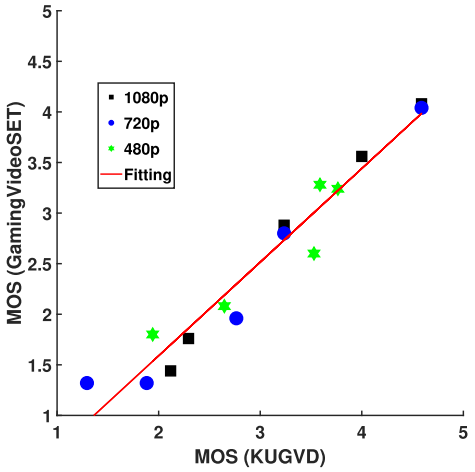

FIGURE 3. Scatter plot of MOS scores and the linear fit corresponding to the anchor conditions (15 conditions of CSGO sequence).

the correlation scores with various metrics. Considering the fact that future work with third-party datasets may not have anchor conditions and that using linear scaling to adjust the MOS scores does not affect the performance of the VQA metrics in terms of their correlation with MOS scores, we finally decided to use the MOS scores from both datasets without any fitting.

As discussed by the authors in [13], since an open dataset is of great use and interest to the research community, we have released the reference and distorted video sequences along with the scores for eight VQA metrics (3 FR, 2 RR and 3 NR) and subjective assessment scores (MOS ratings) as an open source dataset called KUGVD available at https://kingston.box.com/v/KUGVD. Henceforth, we will occasionally use Dataset 1 (D1) and Dataset 2 (D2) to refer to GamingVideoSET and KUGVD respectively.

IV. METHODOLOGY

Fig. 4 shows the methodological framework that is used in this study to develop, test and validate the ML based gaming video quality estimation models. The key blocks of the methodology are feature extraction, feature selection, model development, and performance evaluation and validation. Datasets D1 and D2 were used in the development of NR-GVSQI and NR-GVSQE. For each model, we extracted features and identified the best subset of features to use in model development. After training, validation was performed and each model was further tested on an external dataset which was not used in the model development process. A description of each individual step in the methodological framework is discussed next.

A. FEATURE EXTRACTION

The performance of supervised ML-based predictive models is highly dependent on the features used in model development. Extracting relevant features for supervised learning is therefore critical. Previous statistical analysis has shown that video quality, as perceived by end users, is impacted by the combined influence of many factors such as the initial encoded video quality, content type, and the encoding parameters (e.g., frame rate, RES, BR and the QP) [38], [39]. Besides encoding, which determines the original encoded video quality, network QoS and client-side contexts, such as device type and resolution, further degrade the video quality. Since passive gaming video streaming applications such as Twitch.tv, YouTubeGaming, etc. use HTTP Adaptive Streaming (HAS) technology, which is TCP based, they do not suffer from transmission-related impairments such as packet loss, bit error, etc. Hence, in this work we did not consider the impact of network and user context on the predicted quality. Since our goal is to build a NR model for quality estimation,

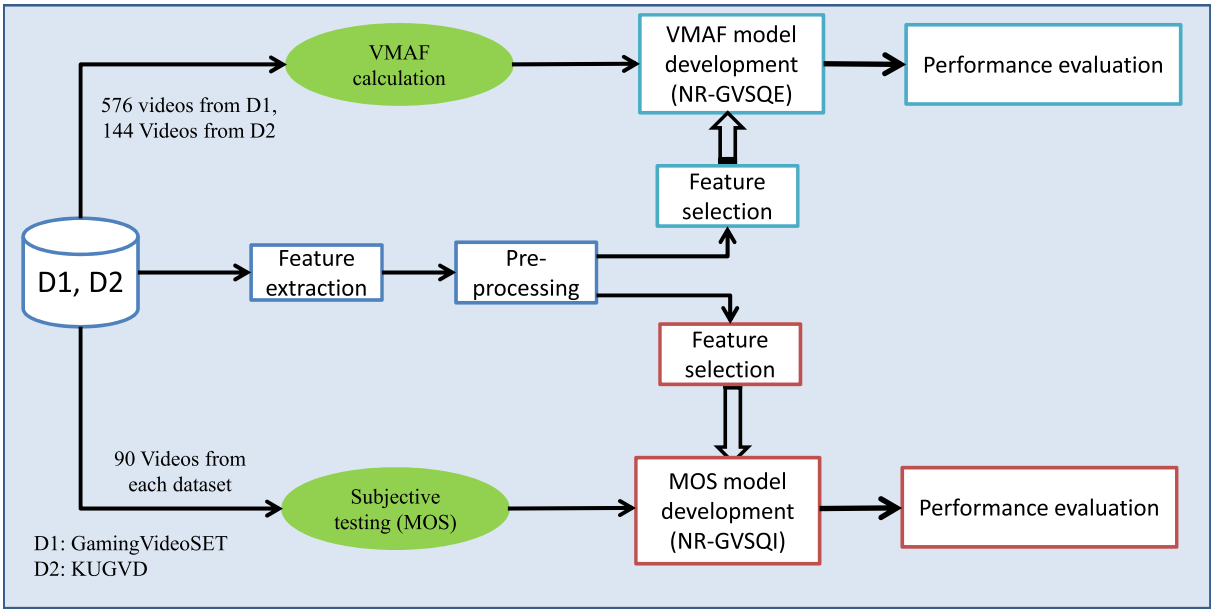

FIGURE 4. Methodological framework used in this work.

TABLE 3. Summary of the fifteen nr features used in this work. the description of some of the features (those extracted using the tool [40]) is based on the description in [41] and [42].

NR Feature	Description
Blockiness	Resulting from the inherent nature of coding algoirthms which operate at the block level. It is one of the most common and visible artifact.
Blockloss	Loss of video data packets possibly during the transmission of the video.
Blur	Reduction of edge sharpness and spatial detail which results in a loss of high frequency information during coding.
Contrast	Difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable.
Exposure	Imbalance of brightness resulting from presence of frames that are too bright or too dark.
Flickering	A visible change in brightness which occurs between the screen refresh events.
Interlacing	Difference between consecutive pixels in columns, interlacing artefact is the result of special video compression where each frame is a connection between two frames in the original video.
Noise	Unwanted, uncontrolled or unprecedented pattern of intensity fluctuations.
Slicing	Artifact due to loss of packets, it occurs when a limited number of video lines is severely damaged.
Spatial Activity	Root mean square over space of the Sobel filtered values of a frame.
Temporal Activity	Root mean square over space of the difference in pixel values of the adjacent frames.
Spatial Information	Maximum over time of the standard deviation over space of the Sobel filtered values of a frame.
Temporal Information	Maximum over time of the standard deviation over space of the difference in pixel values of the adjacent frames.
RES	Resolution of the encoded video.
Bitrate	Number of bits per seconds of the video, reported in this paper in kbps unless stated otherwise.

we extracted features from only the distorted sequences, not relying on any reference information. We extracted 16 NR features for both datasets (GamingVideoSET and KUGVD) based on the encoding process and on content. Hence, each sample of the dataset used in this study is described by 16 NR features based on content (spatial information (SI), temporal information (TI), spatial activity (SA), temporal activity (TA), exposure and contrast) and encoding process (RES, BR, blockiness, blockloss, blur, interlace, noise).

Additionally, we used the output scores of the following three NR metrics as input features for our model development:

1) BIQI, a modular NR metric based on natural scene statistics (NSS).
2) BRISQUE, a NR IQA metric which quantifies the possible loss of naturalness in an image by using the locally normalized luminance coefficients.
3) NIQE, a learning-based NR IQA metric which uses statistical features based on the space domain NSS model.

Table 3 summarizes all the 15 NR features considered in this work. The first eleven NR features were computed using the tool provided by the authors in [40]. The tool provides per-frame scores for each video. We calculated the average of each of these 11 features, which are then used together with the other four features (SI, TI, RES and BR) and three NR metric outputs (which we consider as three features), to select a subset of features that was subsequently used for developing a model that maps these features onto an estimation of the video quality. Although some of these features are related and dependent (e.g., slicing and block loss; exposure and noise metrics; SI and SA; TI and TA), their combinations may result in improved prediction quality, as will be evident later. For a better understanding of these features, we guide the reader to the work in [41] and [42].

In addition, we also use the FR metric VMAF as an estimation of QoE because our earlier works in [7] and [12] have shown that it estimates the subjective quality with high prediction accuracy for gaming videos. It should be noted that due to the reasons mentioned in [12], the three NR metrics were calculated on the downscaled encoded videos, whereas the rest of the features were calculated on upscaled, decoded raw videos.

B. FEATURE SELECTION

The nature of the data used to characterize the relationship between example data and the outcome measure may significantly affect the performance of predictive models. Noisy and unreliable data increase the difficulty of training machine-learning models. Removing redundant features to reduce the dimensionality of the data results in faster and more effective training and reduces the computational costs and the chance of overfitting [43]. The aim of feature selection is to select a subset X_S of the input features, $X = \{x_1, x_2, ..., x_N\}$, so that this subset can predict the outcome measure with a comparable performance with the case when the whole set of features X is used, but with less computational cost [44]. With N features, there are $2^N - 1$ possible subsets of features that should be tested if we want to use exhaustive search.

In order to reduce the complexity, different wrapper and filter feature selection methods [45] can be used to select a subset of the features discussed in Section IV-A. In the forward feature selection method, first the feature with the highest correlation with video quality is selected and progressively more features are added to create a larger subset of features with higher predictive power. Only features that increase the predictive power of the subset are retained. In the backward feature elimination method, all the features are selected as the starting subset and progressively the least promising feature that did not add any predictive power to the

subset of features is eliminated [46]. The steps are repeated until a certain number of features remains or a certain performance level is reached. This reduces the complexity of the feature selection process by reducing the possible number of subsets from $2^N - 1$ to $N(N+1)/2$.

C. MODEL DEVELOPMENT

In this work, we propose two ML based models. The first one aims at estimating the MOS scores obtained via subjective testing, while the second aims at estimating the scores of a well-known objective quality metric (VMAF), that our previous studies identified as the best objective quality metric for gaming video streaming among the state-of-the-art ones analyzed. The first model, presented in Section V, was designed using the MATLAB machine learning toolbox [47] while the second one, presented in Section VI, was designed using Waikato Environment for Knowledge Analysis (WEKA) [48]. As machine learning techniques, we used SVR, Gaussian Process (GP) regression, NN, and Random Forest (RF), which are representative machine learning algorithms that have been used in the domain of video quality prediction and modeling [49]. For an easier understanding of the presented models, we briefly describe the machine learning algorithms used in this work and for a more detailed discussion, we refer the reader to the individual references and to the work of Vega *et al.* [32] where a detailed description of the machine learning algorithms and their application in VQA is presented.

1) Neural Networks [50], commonly referred to as ''artificial'' neural networks is an information processing framework which is modeled after the biological nervous system such as the brain. They usually consist of a large number of interconnected elements (neurons) which work together to solve specific problems. In Section V, we use a two-layer feed-forward network with sigmoid hidden neurons and linear output neurons that is able to fit multi-dimensional mapping problems. The Levenberg-Marquardt backpropagation algorithm is used to train the network. In Section VI, we use MLP which is a class of feed-forward artificial neural network consisting of at least three layers of nodes: an input layer, a hidden layer and an output layer which uses an iterative algorithm based on gradient descent as the backpropagation algorithm for supervised training. Using multiple layers and given the fact that all nodes except the input nodes is a neuron that uses a nonlinear activation function, it is different from a linear perceptron and hence able to distinguish data which is not linearly separable.
2) Support Vector Machines (SVMs) are supervised learning models which use learning algorithms for classification and regression analysis of the input data. Support-vector regression (SVR) is the version of SVM for regression which relies on a kernel function to fit the training data to a function with the error rate within a certain threshold.
3) A Gaussian Process [51] is a stochastic process where any finite subset of the range follows a multivariate Gaussian distribution. Gaussian process regression models are non-parametric kernel based probabilistic models.
4) Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [52]. Random forests are based on the fact that while predictions made by a decision tree may not be accurate, using a combination of them will result in improved prediction accuracy.

The ML techniques used in this work have been selected due to their simplicity: one of the main goals of this work is to propose a light-weight NR metric, which is simple enough to be used in real-world applications. During the design of the metrics we have restricted ourselves to features which are of low complexity and can be extracted in real-time. The same low-complexity criterion is used during the selection of models so that the end metric can be used even on low computational power and energy constrained devices such as smartphones. Our initial work showed promising result with these simple models, hence more complex solutions such as deep learning approaches are not investigated. The major objective of this paper is to investigate different approaches for the selection of appropriate NR features and model design methodologies which can help build a NR quality model with high accuracy of subjective quality prediction specifically for gaming video streaming applications.

V. NO-REFERENCE ESTIMATION OF SUBJECTIVE SCORES (MOS)

A. FEATURE SELECTION AND MODEL DEVELOPMENT

For feature selection we used the backward elimination method explained in Section IV-B for feature selection, and started with all of the 18 features (15 features mentioned in Table 3 and the three NR metrics) discussed in Section IV-A. At the first round of feature selection, to further reduce the complexity, the performance of all the combinations with 17 out of 18 features were examined by SVR, and it was observed that in eight cases the performance of MOS prediction is not significantly worse than the case with all of the 18 features. Then, these eight features were eliminated and the rest of the ten features (TI, RES, BR, BIQI, BRISQUE, NIQE, Blockiness, SA, Blockloss, and TA) remained for further feature reduction.

We also evaluated feature derivation from existing features by PCA method and a mixed method (a combination of the selected features and derived features); however, feature selection using the backward elimination method showed better results. The subset that achieved the best prediction performance was selected for model development. The PLCC score was used to quantify the predictive power of subsets of features.

Table 4 summarizes the correlation scores (PLCC) between the remaining 10 features as well as with MOS

TABLE 4. Correlation (PLCC) between the various features and MOS.

Feature	TI(1)	RES(2)	BR(3)	BIQI(4)	BRISQUE(5)	NIQE(6)	Blockiness (7)	SA(8)	Blockloss(9)	TA(10)	MOS(11)
TI (1)	1.00	-0.02	-0.03	0.32	0.32	0.31	0.22	0.14	0.16	0.85	0.09
RES (2)	-0.02	1.00	0.03	-0.62	-0.61	-0.30	-0.80	0.69	0.67	-0.06	0.04
BR (3)	-0.03	0.03	1.00	0.42	0.45	0.56	0.26	0.24	0.26	-0.12	0.71
BIQI (4)	0.32	-0.62	0.42	1.00	0.87	0.75	0.72	-0.09	-0.18	0.28	0.51
BRISQUE (5)	0.32	-0.61	0.45	0.87	1.00	0.82	0.77	-0.07	-0.06	0.36	0.54
NIQE (6)	0.31	-0.30	0.56	0.75	0.82	1.00	0.70	0.23	0.24	0.28	0.79
Blockiness (7)	0.22	-0.80	0.26	0.72	0.77	0.70	1.00	-0.31	-0.26	0.21	0.42
SA (8)	0.14	0.69	0.24	-0.09	-0.07	0.23	-0.31	1.00	0.79	0.28	0.51
Blockness (9)	0.16	0.67	0.26	-0.18	-0.06	0.24	-0.26	0.79	1.00	0.26	0.43
TA (10)	0.85	-0.06	-0.12	0.28	0.36	0.28	0.21	0.28	0.26	1.00	0.08
MOS (11)	0.09	0.04	0.71	0.51	0.54	0.79	0.42	0.51	0.43	0.08	1.00

TABLE 5. Performance of various feature subsets for different training and test dataset combinations. the best performing cases are shown in bold.

Predictors (features)	Training			Test		
	MSE	PLCC (%)	SROCC (%)	MSE	PLCC (%)	SROCC (%)
1,2,3,4,5,6,7,8,9,10 (10F)						
Trained on D1, Tested on D2	0.10	0.93	0.94	0.58	0.84	0.85
Trained on D2, Tested on D1	0.05	0.98	0.98	0.22	0.87	0.86
1,2,3,4,5,7,8,9,10 (9F)						
Trained on D1, Tested on D2	0.11	0.93	0.93	0.42	0.85	0.86
Trained on D2, Tested on D1	0.07	0.97	0.97	0.22	0.86	0.86
1,2,3,4,5,7,8,10 (8F)						
Trained on D1, Tested on D2	0.12	0.92	0.93	0.42	0.85	0.85
Trained on D2, Tested on D1	0.07	0.97	0.97	0.22	0.87	0.86
1,2,3,4,5,8,10 (7F)						
Trained on D1, Tested on D2	0.14	0.91	0.91	**0.32**	**0.89**	**0.89**
Trained on D2, Tested on D1	0.08	0.96	0.96	**0.22**	**0.87**	**0.86**
1,2,4,5,8,10 (6F)						
Trained on D1, Tested on D2	0.22	0.86	0.85	0.83	0.74	0.75
Trained on D2, Tested on D1	0.13	0.94	0.94	0.99	0.73	0.73
1,2,4,8,10 (5F)						
Trained on D1, Tested on D2	0.20	0.87	0.86	0.94	0.70	0.72
Trained on D2, Tested on D1	0.21	0.90	0.91	0.41	0.73	0.72
1,4,8,10 (4F)						
Trained on D1, Tested on D2	0.22	0.85	0.85	0.80	0.71	0.73
Trained on D2, Tested on D1	0.26	0.88	0.88	0.42	0.72	0.72

Features selected are: TI(1), RES(2), BR(3), BIQI(4), BRISQUE(5), NIQE(6), Blockiness(7), SA(8), Blockloss(9), TA(10)
D1: GamingVideoSET; D2: KUGVD
Results are averaged over 100 iterations

scores (GamingVideoSET and KUGVD combined). For the feature selection part, both datasets (180 samples) were used, and then the model was trained and validated on GamingVideoSET, and its scalability (generalization) tested on KUGVD (and vice versa). The anchor conditions were used only during the training phase and were removed during the testing phase (since we had 15 anchor conditions in total, we had 90 samples for training and 75 samples for testing). Since TI and TA are calculated almost identically (see Table 3), it can be seen that they have a high correlation score of 0.85. As the videos are encoded at different resolutions, there is a high correlation between blockiness and RES, as expected. Also, since all three NR metrics are based on Natural Scene Statistics (NSS), they have a high correlation among themselves. Also, Blockloss is found to have a high correlation with SA.

We evaluated the performance of both SVR and NN with all feature subset combinations for the two different training and test dataset combinations. Table 5 shows the performance results in terms of Mean Square Error (MSE), PLCC and SROCC scores of the best performing algorithm (NN) for various feature subsets. Here we have used the backward elimination method to reduce the number of features from 10 to 4 in 6 consecutive steps. It can be observed that for all feature subsets and training and test dataset combinations, NN performs better than SVR. For NN regression, the number

of layers is 2 and the number of hidden neurons is 10. Levenberg-Marquardt is chosen as the training algorithm. In order to avoid over-fitting, the error is checked on validation; if it keeps increasing beyond a fixed known limit (set here to 6), then the training will stop. The result of the trained network at the point of increasing error on validation is used for the test.

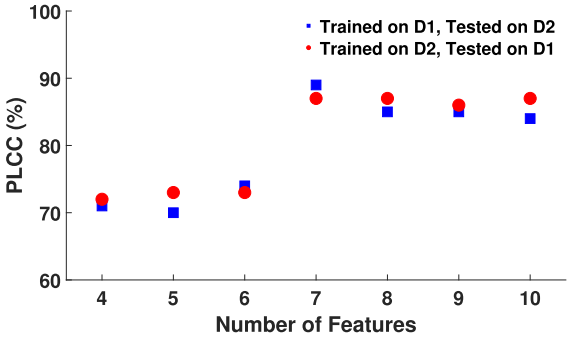

FIGURE 5. PLCC (%) variation with different number of features, for different training and test scenarios.

Fig. 5 shows the performance of the model in terms of PLCC scores with respect to the different number of features, considering two different training and test dataset combination scenarios as follows:

- *'Trained on D1, Tested on D2'*: Under test scenario, D1 and D2 were used as the ''training'' and ''test'' parts respectively consisting of 90 and 75 samples, respectively. This was repeated for 100 iterations for each feature subset combination.
- *'Trained on D2, Tested on D1'*: Same as *'Trained on D1, Tested on D2'* scenario but with D1 and D2 interchanged.

It can be observed that for both scenarios the prediction accuracy increases when the number of features is reduced from ten to seven. Further reduction in the number of features then reduces the prediction accuracy. The scalability (generalization) of the model is really tested when different datasets are used for training and testing. This is not considered in many research methodologies in QoE modeling and has resulted in optimistic performance results. Based on the results presented in Table 5 for the two different training and test scenarios, we consider that the NN model using 7 features results in the optimal performance which we refer to as NR-GVSQI.

Based on the results presented in Table 5 and Fig. 5, it can be observed that the performance results do not vary much when different datasets are used for training and testing (*'Trained on D1, Tested on D2'*, *'Trained on D2, Tested on D1'*) which leads us to the following conclusions:

- The proposed model, NR-GVSQI trained on a gaming video of a particular gameplay scene from a game is robust enough to predict the quality of another gaming video of another gameplay scene of the same game.
- When trained on a set of gaming videos from one dataset, NR-GVSQI is robust enough to predict the quality of a gaming video from a totally new game belonging to a different genre and complexity (both datasets consist of a game from a totally different genre; Project Cars in D1 and Overwatch in D2, see Fig. 2).

TABLE 6. VQA metrics performance on the two datasets. the proposed metric's performance is shown in bold.

		GamingVideoSET	KUGVD
FR Metrics	PSNR	0.74	0.80
	SSIM	0.80	0.89
	VMAF	0.87	0.92
RR Metrics	STRREDopt	-0.71	-0.73
	SPEEDQA	-0.71	-0.70
NR Metrics	BRISQUE	-0.49	-0.62
	BIQI	-0.43	-0.60
	NIQE	-0.77	-0.85
	MEON	-0.35	-0.43
	NR-GVSQI	**0.87**	**0.89**

Table 6 shows the performance of nine state-of-the-art VQA metrics on the two datasets along with the performance of the proposed metric, NR-GVSQI. In addition to the three NR metrics considered during model development, we also compare the performance of our proposed metric with the recently proposed deep NN based metric MEON discussed in Section II which has been shown to outperform existing learning based as well as traditional NR metrics. Due to the use of proprietary SSIMplus in V-MEON, its the model implementation is no longer available publicly and hence could not be evaluated on our datasets. For the evaluation of MEON we have used the implementation and default settings as provided by the authors in the respective publication. The scores were computed on a per-frame basis and averaged over the whole video to obtain a final score. Due to non-availability of ground truth scores for our datasets, the model could not be re-trained and was evaluated using the trained weights provided by the authors.

It can be observed that the proposed metric NR-GVSQI results in the best performance on both datasets when compared to the four NR metrics. Considering GamingVideoSET, the proposed metric NR-GVSQI achieves a correlation of 0.87, which is almost the same as that achieved by the state-of-the-art FR metric, VMAF. For KUGVD, the metric achieves almost similar performance to SSIM, a widely used FR VQA metric. Comparing the performance across both datasets, it can be observed that while the performance of the state-of-art NR metric NIQE varies quite a lot, the performance achieved by NR-GVSQI is more stable across the two datasets. Among the four NR metrics, MEON results in the worst performance across both datasets (considering all 90 stimuli each) which is surprising given its high performance on different IQA datasets. This indicates that a machine learning based NR metric designed and tested on non-gaming videos does not necessarily perform well on gaming datasets and vice versa, hence establishing the need for customized NR metrics for gaming videos. Given the limited amount of training data we had, we expect that the proposed metric NR-GVSQI, when trained on a much

bigger dataset, will result in an improved and more stable performance across different gaming videos and hence will be more generalizable to be used for quality estimation of gaming video streaming applications.

B. DISCUSSION ON FEATURE SELECTION

Based on the correlation values presented earlier in Table 4 it can be observed that the correlation between the features BIQI (4) and SA (8) with MOS is 0.51 and the correlation between the TI (1) and TA (10) with MOS is 0.09 and 0.08, respectively, which is very low. This implies that the selected features do not necessarily have a high correlation with the predicted entity. More interestingly, if we consider the correlation between the selected features, we can see for example that the correlation between TI (1) and TA (10) is 0.85, which as expected, is quite high. Hence, if feature selection would have been performed just based on the correlation values in Table 4, feature TI(1) and TA(10) both would not have been included in the same feature subset. Therefore, it can be concluded that the prediction performance of the features when considered individually and in a group can vary a lot. Feature selection using just domain knowledge or based on correlation with the prediction entity, as we saw here, is not enough and needs to be supplemented by feature selection methods such as backward elimination method as used in this work.

VI. NO-REFERENCE ESTIMATION OF FULL REFERENCE OBJECTIVE METRIC (VMAF)

In the previous section, we proposed a machine learning based no reference model which was trained and evaluated in terms of its capability to estimate MOS. The limit of this approach is that the training and test data available consist of only 165 stimuli (90 stimuli from each dataset minus the 15 common conditions). Yet, conducting more subjective tests is time consuming, expensive and impractical, especially when creating a large dataset consisting of a large number of videos and encompassing various distortions. Hence, in this section we explore the possibility of developing a ML-based model to predict, rather than MOS, the best performing objective VQA score (still using NR features); a dataset covering a wide range of content and conditions is more practical and easier to create with VQA scores. We use the GamingVideoSET, which consists of a total of 576 distorted sequences, for model development and 120 sequences from KUGVD (excluding the 24 anchor conditions) for testing purposes. Since VMAF was found to have a very high correlation with MOS for both datasets [7], [12], we calculated for both datasets the VMAF scores, which are then used as the ground truth for the ML algorithms. Fig. 6 shows that the distribution of encoded video quality - in terms of VMAF - is well spread from low to high.

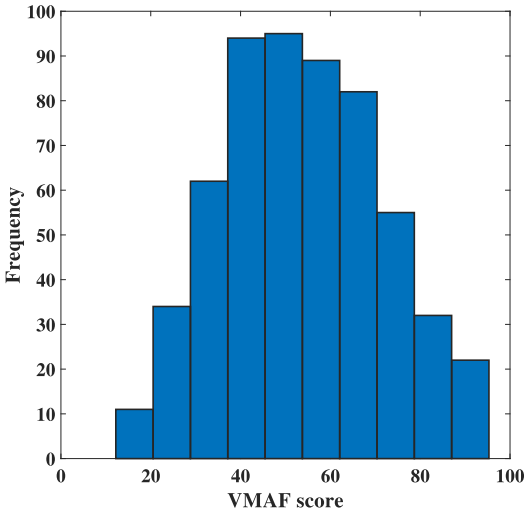

FIGURE 6. Histogram showing the distribution of the quality (VMAF) of the encoded video sequences from gaming video SET used for training of the model.

A. FEATURE REDUCTION AND MODEL DEVELOPMENT

The Waikato Environment for Knowledge Analysis was adopted for feature reduction and model design for this metric. Based on our domain knowledge, as well as based on the results from our work in the previous section, we used 14 candidate predictors from the encoding process (BR, RES), content (SI, TI, SA, TA, Noise, Exposure and Contrast), compression artifacts (blur and blockiness) and no reference video quality metrics (BIQI, BRISQUE and NIQE) as the initial feature set. Four features considered for the earlier work (Blockloss, Interlacing, Flickering and Slicing) were not considered as they are not valid for the encoding conditions considered in the datasets used in this work.

Table 7 shows the correlation of features with VMAF and with different features in terms of PLCC scores. The features with the best correlation with VMAF are BR (0.68), SA (0.65), BLUR (−0.62), RES (0.55) and SI (0.51). No single feature is robust enough to predict the encoded video quality with acceptable accuracy. However, the correlation between BR with SA, BLUR, RES and SI is relatively low. A combination of BR and the three quality metrics may not yield high predictive power given that the correlation between BR and BIQI, BRISQUE and NIQE is relatively high. Combining BIQI, BRISQUE and NIQE to predict VMAF may not yield high prediction value due to the high inter-correlation between them, but as observed in results from model design in Section V, this may not always be the case. Hence, as done previously in Section V, it is necessary to conduct a feature selection process to determine a subset of features from the considered initial 14 features. Towards this end, we extracted the above mentioned initial 14 features for the full GamingVideoSET and KUGVD datasets using the same approach as was used previously. For model design and validation we used the extracted features from GamingVideoSET for feature selection purposes to determine subsets of features for model development. We used the WEKA correlation based feature selection (CFS) function to select subsets of features [53]. This allowed us to reduce data dimensionality without having to manually evaluate all

TABLE 7. Correlation (PLCC) between the various features considering all 576 sequences from GamingVideoSET.

Feature	RES	BR	SI	TI	BIQI	BRISQUE	NIQE	Blockiness	SA	TA	Blur	Noise	Exposure	Contrast	VMAF
RES	1.00	0.14	0.76	0.02	0.64	0.60	0.35	-0.82	0.77	0.06	-0.91	0.08	0.02	0.05	0.55
BR	0.14	1.00	0.23	0.04	-0.30	-0.36	-0.56	0.17	0.27	0.11	-0.21	0.46	0.03	0.02	0.68
SI	0.76	0.23	1.00	0.21	0.40	0.34	0.12	-0.46	0.92	0.27	-0.62	0.14	0.45	0.38	0.51
TI	0.02	0.04	0.21	1.00	0.32	0.27	0.24	-0.11	-0.06	0.86	0.10	-0.13	0.29	0.58	-0.20
BIQI	0.64	-0.30	0.40	0.32	1.00	0.89	0.77	-0.72	0.27	0.32	-0.55	-0.34	0.01	0.13	-0.04
BRISQUE	0.60	-0.36	0.34	0.27	0.89	1.00	0.82	-0.69	0.22	0.29	-0.50	-0.33	0.02	0.10	-0.11
NIQE	0.35	-0.56	0.12	0.24	0.77	0.82	1.00	-0.59	0.01	0.18	-0.23	-0.35	-0.04	0.02	-0.41
Blockiness	-0.82	0.17	-0.46	-0.11	-0.72	-0.69	-0.59	1.00	-0.43	-0.11	0.67	0.04	0.01	-0.07	-0.14
SA	0.77	0.27	0.92	-0.06	0.27	0.22	0.01	-0.43	1.00	-0.06	-0.69	0.22	0.32	0.22	0.65
TA	0.06	0.11	0.27	0.86	0.32	0.29	0.18	-0.11	-0.06	1.00	0.10	-0.08	0.46	0.61	-0.19
Blur	-0.91	-0.21	-0.62	0.10	-0.55	-0.50	-0.23	0.67	-0.69	0.10	1.00	-0.11	0.20	0.22	-0.62
Noise	0.08	0.46	0.14	-0.13	-0.34	-0.33	-0.35	0.04	0.22	-0.08	-0.11	1.00	0.05	-0.14	0.28
Exposure	0.02	0.03	0.45	0.29	0.01	0.02	-0.04	0.01	0.32	0.46	0.20	0.05	1.00	0.77	-0.05
Contrast	0.05	0.02	0.38	0.58	0.13	0.10	0.02	-0.07	0.22	0.61	0.22	-0.14	0.77	1.00	-0.04
VMAF	0.55	0.68	0.51	-0.20	-0.04	-0.11	-0.41	-0.14	0.65	-0.19	-0.62	0.28	-0.05	-0.04	1.00

TABLE 8. Results of model development for different sub-features using GP, MLP, SVR and RFML algorithms. The best performing result is shown in bold italics.

Features	GP			MLP			SVR			RF		
	PLCC	MAE	RMSE	PLCC	MAE	RMSE	PLCC	MAE	RMSE	PLCC	MAE	RMSE
1	0.77	9.04	11.17	0.64	11.22	14.01	0.67	10.58	13.16	0.79	8.74	10.90
2	0.87	7.18	8.64	0.76	9.60	11.82	0.87	6.97	8.76	0.85	7.38	9.39
3	0.89	6.51	8.07	0.80	9.01	11.49	0.88	6.43	8.41	0.87	6.72	8.65
4	0.93	4.97	6.40	0.88	7.12	9.09	0.95	4.18	5.77	0.93	4.79	6.29
5	0.94	4.77	6.19	0.91	6.15	7.82	0.96	3.45	4.91	0.95	4.52	5.82
6	0.96	4.16	5.24	0.92	5.83	7.28	0.99	2.06	2.92	0.96	4.05	5.10
7	***0.96***	***3.98***	***5.10***	***0.93***	***5.11***	***6.54***	***0.99***	***1.74***	***2.44***	***0.96***	***4.03***	***5.07***
8	0.96	3.83	4.91	0.93	5.13	6.46	0.99	1.35	2.02	0.97	3.74	4.71
9	0.97	3.73	4.78	0.96	4.21	5.25	0.99	1.36	2.00	0.97	3.77	4.71
10	0.97	3.74	4.80	0.95	4.74	5.99	0.99	1.24	1.81	0.97	3.77	4.73
11	0.97	3.34	4.24	0.98	3.16	3.97	1.00	1.11	1.68	0.97	3.38	4.34
12	0.98	3.23	4.11	0.98	2.72	3.40	1.00	1.05	1.60	0.98	3.24	4.15
13	0.98	3.18	4.08	0.98	2.53	3.10	1.00	1.07	1.61	0.97	3.44	4.34
14	0.98	3.19	4.09	0.98	2.75	3.48	1.00	1.05	1.60	0.97	3.36	4.26

possible combinations of features. CFS evaluates the predictive worth of a subset of features by considering the predictive power of each feature together with the amount of redundancy between features. Preference is given to subsets that are highly correlated with VMAF whilst also having a low correlation between them. We selected subsets of features with the number of features in each subset increasing from 1 to 14 features. While this might not always be the optimal method for testing different feature set combinations, it works with reasonable accuracy considering our feature subset and model design as will be shown later. Each subset was then used as variables to develop four regression models using four different ML-based algorithms (GP, MLP, SVR and RF). These are some of the popular and frequently used ML algorithms in image/video quality prediction [27].

Using the extracted features from GamingVideoSET, we developed prediction models using the 10-fold cross validation methodology. The data were randomly divided into 10 sub-datasets. Nine sub-datasets were used for training a machine learning model and one was left out to test the model. This is repeated ten times until all sub-datasets have been used for training and testing. This methodology has the advantage that all data are used for training and testing. It is commonly used in machine learning to avoid overfitting [54]–[56], as was also used in the previous section. The performance of each model was averaged over the testing processes in order to determine the general performance.

Table 8 shows the performance of each subset of features in predicting video quality in terms of PLCC, mean absolute error (MAE) and RMSE. Increasing the number of features increases the prediction accuracy for all ML algorithms. However, there is an optimum number of features that provides an optimum balance between accuracy and complexity in terms of the number of features needed to achieve acceptable performance. Increasing the number of features beyond this number minimally improves the performance. However, this improvement is at the expense of increased complexity and increased computational requirement, which may limit usability in real time especially for thin clients.

For example, increasing the number of features used in the SVR model from 3 to 7 features increases PLCC by 12.5%. Yet doubling the number of features from 7 to 14 for the same algorithm increases PLCC by only 0.6%. Doubling the number of features does not improve the performance significantly, and only increases the model complexity (mainly due to increased feature extraction tasks as well as model

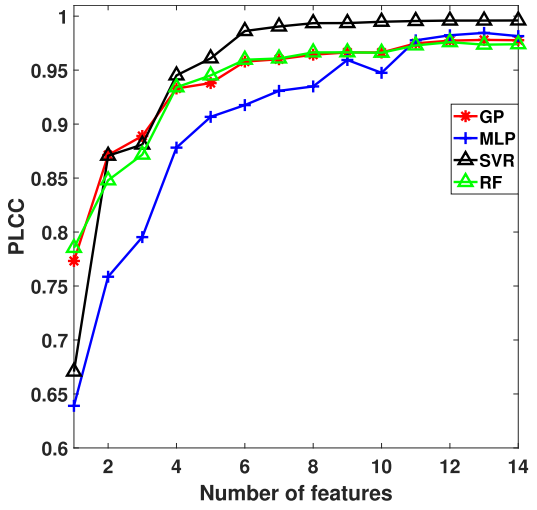

FIGURE 7. Impact of number of features on prediction accuracy for GP, MLP, SVR, and RF learning algorithms.

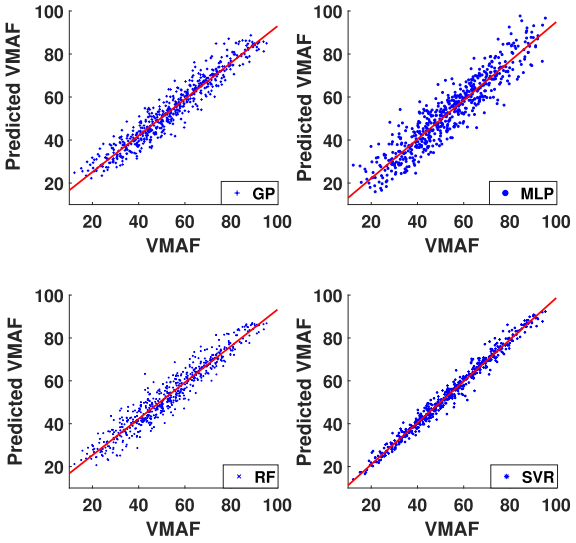

FIGURE 8. Prediction performance of GP, MLP, RF and SVR prediction models on the training dataset (GamingVideoSET).

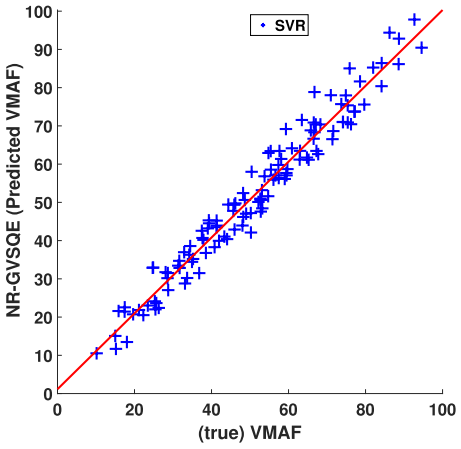

FIGURE 9. NR-GVSQE (Predicted VMAF) scores vs. (true) VMAF scores considering KUGVD dataset.

computation time). Fig. 7 shows the relationship between the number of features and the performance for the four learning algorithms. The results clearly show saturation in performance with an increased number of features for all algorithms. SVR obtained superior performance over all the other algorithms, with 7 features being optimal (*RES, BR, SI, TI, Contrast, Blur and Exposure*). We refer to this prediction model as NR-GVSQE. This is similar to our observation reported in Section V during our NR-GVSQI model design and evaluation using MOS ratings as the labels where also 7 features resulted in the optimum prediction model. The subset of features is different from those used in NR-GVSQI, which is not surprising as the relationship between the various features and MOS is not exactly the same as for VMAF. Henceforth, the VMAF scores as obtained from the distorted video sequences will be referred to as *(true) VMAF* while the VMAF scores as predicted using NR-GVSQE will be referred to as *predicted VMAF*.

Fig. 8 shows the performance of the models in terms of prediction accuracy during model development. The figure clearly shows that the quality prediction model based on SVR is of superior quality. We, therefore, selected this model for testing using the KUGVD dataset.

The selected SVR model was inherently validated during development due to the nature of 10-fold cross validation methodology. In practice, this is not usually enough and the final testing is usually conducted on an external dataset that was not used in model development. We externally validated and tested the model on KUGVD which is an unseen dataset and was not used in model development. The dataset has 120 samples (excluding the anchor video conditions) and has the same features as the training dataset. Fig. 9 shows the predicted VMAF by the model plotted against the (true) VMAF of KUGVD. The predicted VMAF is highly correlated with (true) VMAF, with a PLCC of 0.98.

Table 9 presents a comparative performance evaluation of our proposed metric versus popular FR (PSNR, SSIM), RR (STRREDopt, SpEED-QA) and NR (BIQI, BRISQUE, and NIQE) metrics to predict the VMAF of the KUGVD dataset. It can be observed that the proposed model outperforms these quality metrics by a huge margin in terms of correlation with VMAF. Since the ultimate goal of any IQA/VQA metric is to be able to predict the subjective quality as presented in Section V, we evaluate the performance of the model presented here - developed based on VMAF scores - in terms of correlation with respect to MOS scores from the subjective dataset.

Table 10 compares the performance of (true) VMAF and predicted VMAF scores vs. MOS on the KUGVD dataset, considering 75 stimuli (excluding anchor conditions). It can be observed that our proposed model, which was trained using VMAF scores from GamingVideoSET, when tested on

TABLE 9. Correlation (PLCC) of various VQA metrics and the proposed model, NR-GVSQE, W.R.T VMAF scores for KUGVD dataset. The best performing model is shown in bold.

Method	PLCC
PSNR	0.89
SSIM	0.84
STRREDopt	-0.66
SpeedQA	-0.64
BIQI	-0.35
BRISQUE	-0.4
NIQE	-0.74
NR-GVSQE	**0.97**

TABLE 10. Performance evaluation in terms of PLCC and SROCC of (true) VMAF and NR-GVSQE (PREDICTED VMAF) scores W.R.T MOS scores from KUGVD dataset (excluding 15 anchor conditions).

Metric	PLCC	SROCC
(true) VMAF	0.930	0.934
NR-GVSQE	0.905	0.913

an unknown dataset results in similar performance as (true) VMAF scores with respect to MOS ratings. It is important to note here that our trained model utilizes NR features and hence is a NR metric, compared to VMAF which is a FR metric.

Compared to NR-GVSQI which was trained on MOS scores, the performance of NR-GVSQE is approximately 1.5% better on KUGVD in terms of PLCC scores. The improved performance of NR-GVSQE compared to NR-GVSQI can be attributed to the fact the model design was performed using a much larger dataset due to the availability of objective VQA scores. The gain of 1.5% might not look a major improvement, considering the fact that both NR-GVSQI and NR-GVSQE use seven input features. It must, however, be noted that while NR-GVSQI uses the NR metrics BRISQUE and BIQI as input features, NR-GVSQE does not use any NR metric scores and uses only very basic NR features such as contrast, blur, and exposure along with basic features such as resolution, bitrate, SI and TI values. Hence, NR-GVSQE is of much lower complexity as compared to NR-GVSQI, with the added advantage that such model design can be performed on a huge dataset with multiple distortion artefacts without the need for any subjective (MOS) ratings.

VII. DISCUSSION

We will discuss in this Section the specificity of the model for gaming video and the comparison with other gaming video quality models.

The model design and performance on the gaming video datasets benefit from the inherent characteristics of the gaming videos (less variation in SI due to repetitive game elements [57], a difference in subjective opinions, etc. [7], which is not true for ordinary videos). For example, as discussed in [57], video games have special content characteristics in that they share the spatial and temporal features between different scenes of the same game. In fact, each game has a special motion pattern and a quite constant spatial complexity, as games are made of a pool of reused objects, which can be exploited by the machine learning algorithms, with possible increased performance for such gaming videos. In light of these factors, we argue that the while the proposed models are shown to work with high accuracy on the gaming video datasets considered in this work, it does not necessarily hold true for other non-gaming datasets (currently an ongoing work).

As discussed before, gaming video streaming applications have so far not gained much attention from the research community. So far, in parallel to our work, there are two similar works carried out by the authors in [58] and [59] who also proposes machine learning based NR models: NR-GVQM and nofu, respectively.

NR-GVQM [58] is a SVR based model with Gaussian kernel which uses nine frame level input features and is trained and validated on an open-source gaming video dataset, GamingVideoSET (see Section III). The model is trained using per-frame scores from 408 distorted video sequences (369000 frames) using VMAF scores as the target output, similar to the approach used in our proposed model NR-GVSQE (see Section VI). The model, when tested on the rest 144 distorted sequences, resulted in a correlation score of 0.98 with VMAF, while our proposed model NR-GVSQE achieves a correlation of 0.97 with VMAF. On the subjective dataset (90 video sequences), the model achieves a correlation of 0.89 with MOS. Compared to NR-GVQM, our model, NR-GVSQE achieves a higher correlation of 0.905 with MOS on an unknown dataset (KUGVD) using a lower number of features (seven compared to nine) and is of much lower complexity, as NR-GVSQE uses input features per video unlike NR-GVQM which uses per-frame scores for the final quality prediction.

Nofu [59] considers 12 different NR feature values per frame which are then divided into three equidistant groups, independent of the duration of the video. For each group, three values for each feature - the first value, the mean and standard deviation for each group is calculated, which results in a total of nine values per feature and a total of 108 pooled features values (considering the 12 selected features). The features are extracted from 360p center crop of the rescaled input video (irrespective of the native video resolution) after which the ExtraTreeRegressor method is used for feature selection using $0.5 \times mean$ as the threshold value and Random Forest as the choice of their regression algorithm. Similar to the aforementioned model, this model also uses VMAF scores (rescaled to 1-5) as target output. The model, when trained and tested on the GamingVideoSET via 10-fold cross validation, is shown to achieve a correlation of 0.96 as compared to 0.97 for our proposed model, NR-GVSQE. The proposed model, when tested on the subjective dataset part of GamingVideoSET (90 videos) via 10-fold cross validation, is shown to achieve a correlation of 0.91. In addition,

the authors performed a source video based train and validation fold approach for subjective score prediction. For the 6 different video sources, they use 5 sources for training and 1 for validation, for which they achieved a correlation of 0.77. As discussed earlier, such an evaluation is hard because of the fact that each gaming video is from a different gaming genre and hence such an evaluation of a metric when tested on an unknown video(s) offers a more critical evaluation of the proposed model's performance for real-world applications. In contrast, our proposed model NR-GVSQE when trained on GamingVideoSET and tested on KUGVD containing different videos from the same as well as different games, achieves a correlation of 0.905.

Although both NR-GVQM and nofu appear to be promising models, due to lack of a second test dataset for the evaluation of the model performance, as discussed above, the actual performance of the models for real-world applications is not established. This also establishes the necessity of another open-source gaming dataset, such as KUGVD as presented in this paper, which can be used for proper validation of proposed models for gaming streaming applications.

VIII. CONCLUSION

Subjective quality assessment of encoded gaming video is a necessity, yet it is time consuming, expensive, and not applicable in real time quality assessment scenarios. As a consequence, the development of objective quality assessment metrics is necessary. For some applications, such as passive gaming video streaming, FR and RR metrics are not suitable due to the unavailability of source information. On the other side, it has been shown that No-Reference (NR) quality metrics developed for natural video content are not suitable for compressed gaming video. Towards this end, we presented in this paper two machine learning based NR metrics, NR-GVSQI and NR-GVSQE for gaming video quality prediction. Our proposed models, which are designed using supervised learning algorithms using MOS and FR Metric (VMAF) scores as the target output, are shown to perform better than the current state-of-the-art NR metrics, in the latter case achieving performance close to the state-of-the-art FR metric (VMAF). One of the major advantages of the proposed models is that they use a small number of features which can be extracted in real-time, hence the models can be used for real-time quality estimation of encoded gaming videos for live gaming video streaming applications.

Due to the inherent nature of the available datasets, the proposed models are limited to only compression and scaling artefacts. Also, currently both datasets are limited in scope considering the number of different games and the resolution-bitrate pairs considered. Since the datasets consist of videos compressed with the H.264 encoder, their performance on videos encoded with other newer encoders such as HEVC, VP9, or AV1 is an open question which we plan to explore in our future work, along with the creation of open-source datasets with an increased variety of games.

REFERENCES

[1] D. Fitzgerald and D. Wakabayashi, "Apple quietly builds new networks," *Wall Street J.*, Feb. 2014. Accessed: Jan. 7, 2019. [Online]. Available: https://www.wsj.com/articles/apple-quietly-builds-new-networks-1391474149

[2] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-T Rec. BT.500-13, ITU-T Recommendation, Jan. 2012.

[3] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[5] Netflix. (Jun. 2016). *Toward A Practical Perceptual Video Quality Metric*. Accessed: Jan. 15, 2019. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[6] *User Requirements for Objective Perceptual Video Quality Measurements in Digital Cable Television*, document ITU-T Rec. J.143, ITU-T Recommendation, Apr. 2008.

[7] N. Barman, S. Zadtootaghaj, M. G. Martini, and S. Möller, and S. Lee, "A comparative quality assessment study for gaming and non-gaming videos," in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–6.

[8] N. Barman and M. G. Martini, "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 codec comparison for live gaming video streaming," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Erfurt, Germany, May 2017, pp. 1–6.

[9] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[10] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[11] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[12] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An evaluation of video quality assessment metrics for passive gaming video streaming," in *Proc. 23rd Packet Video Workshop (PV)*, Amsterdam, The Netherlands, 2018, pp. 7–12.

[13] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "GamingVideoSET: A dataset for gaming video streaming applications," in *Proc. 16th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Amsterdam, The Netherlands, Jun. 2018, pp. 1–6.

[14] M. Slanina and V. Ricny, "Estimating PSNR without reference for real H.264/AVC sequence intra frames," in *Proc. 18th Int. Conf. Radioelektronika*, Prague, Czech Republic, Apr. 2008, pp. 1–4.

[15] B. Girod, "What's wrong with mean-squared error?" in *Digital Images and Human Vision*, 1993, pp. 207–220. [Online]. Available: http://dl.acm.org/citation.cfm?id=197765.197784

[16] X. Jiang, F. Meng, J. Xu, and W. Zhou, "No-reference perceptual video quality measurement for high definition videos based on an artificial neural network," in *Proc. Int. Conf. Comput. Electr. Eng.*, Phuket, Thailand, Dec. 2008, pp. 424–427.

[17] A. Khan, L. Sun, and E. Ifeachor, "Learning models for video quality prediction over wireless local area network and universal mobile telecommunication system networks," *IET Commun.*, vol. 4, no. 12, pp. 1389–1403, Aug. 2010.

[18] M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor," in *Proc. 4th Int. Congr. Image Signal Process.*, Shanghai, China, vol. 1, Oct. 2011, pp. 517–521.

[19] C. Wang, X. Jiang, F. Meng, and Y. Wang, "Quality assessment for MPEG-2 video streams using a neural network model," in *Proc. IEEE 13th Int. Conf. Commun. Technol.*, Jinan, China, Sep. 2011, pp. 868–872.

[20] W. Cherif, A. Ksentini, and D. Négru, "No-reference Quality of Experience estimation of H264/SVC stream," in *Proc. IEEE Globecom Workshops*, Anaheim, CA, USA, Dec. 2012, pp. 1346–1351.

[21] H. E. Khattabi, D. Aboutajdine, and A. Tamtaoui, "Predict the MOS and the PSNR by the neural network," in *Proc. Int. Conf. Multimedia Comput. Syst.*, Tangier, Morocco, May 2012, pp. 418–421.

[22] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," in *Proc. IEEE Consumer Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2012, pp. 127–131
[23] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.
[24] S. L. Sunala and P. R. Anurenjan, "A novel video quality measurement using ANN," in *Proc. Annu. Int. Conf. Emerg. Res. Areas Int. Conf. Microelectron., Commun. Renew. Energy (AICERA/ICMiCR)*, Kanjirapally, India, Jun. 2013, pp. 1–4.
[25] Y. Kang, H. Chen, and L. Xie, "An artificial-neural-network-based QoE estimation model for video streaming over wireless networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Xi'an, China, Aug. 2013, pp. 264–269.
[26] Y. Xue, B. Erkin, and Y. Wang, "A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 134–139, Jan. 2015.
[27] M. T. Vega, E. Giordano, D. C. Mocanu, D. Tjondronegoro, and A. Liotta, "Cognitive no-reference video quality assessment for mobile streaming services," in *Proc. 7th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Pylos-Nestoras, Greece, May 2015, pp. 1–6.
[28] K. Zheng, X. Zhang, Q. Zheng, W. Xiang, and L. Hanzo, "Quality-of-experience assessment and its application to video services in LTE networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 70–78, Feb. 2015.
[29] M. Juayek and R. Sotelo, "An artificial neural network approach for no-reference high definition video quality assessment," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Nara, Japan, Jun. 2016, pp. 1–3.
[30] J. Guo, K. Zheng, G. Hu, and L. Huang, "Packet layer model of HEVC wireless video quality assessment," in *Proc. 11th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Nagoya, Japan, Aug. 2016, pp. 712–717.
[31] C. Wang, L. Su, and Q. Huang, "CNN-MR for no reference video quality assessment," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Changsha, China, Jul. 2017, pp. 224–228.
[32] M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, "Deep learning for quality assessment in live video streaming," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 736–740, Jun. 2017.
[33] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[34] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[35] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. 26th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2018, pp. 546–554. doi: 10.1145/3240508.3240643.
[36] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document P.910:ITU-T Rec, ITU-T Recommendation, Apr. 2008.
[37] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pëpion, and P. Le-Callet, "Aligning subjective tests using a low cost common set," Euro ITV, Lisbon, Portugal, Jun. 2011.
[38] L. Anegekuh, L. Sun, and E. Ifeachor, "End to end video quality prediction for HEVC video streaming over packet networks," in *Proc. 1st Int. Conf. Commun., Signal Process., Their Appl. (ICCSPA)*, Sharjah, UAE, Feb. 2013, pp. 1–6.
[39] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 258–262, Jun. 2011.
[40] *AGH Video Quality Indicators*. Accessed: Jan. 15, 2018. [Online]. Available: http://vq.kt.agh.edu.pl/metrics.html
[41] M. Leszczuk, M. Hanusiak, I. Blanco, A. Dziech, J. Derkacz, E.Wyckens, and S. Borer, "Key indicators for monitoring of audiovisual quality," in *Proc. 22nd Signal Process. Commun. Appl. Conf. (SIU)*, Trabzon, Turkey, Apr. 2014, pp. 2301–2305.
[42] M. Leszczuk, M. Hanusiak, M. C. Q. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators," *Multimedia Tools Appl.*, vol. 75, no. 17, pp. 10745–10767, Sep. 2016.
[43] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, vol. 454. Springer, 2012.
[44] I. Witten, E. Frank, and M. Hall, "Data mining: Practical machine learning tools and techniques," in *The Morgan Kaufmann Series in Data Management Systems*. Amsterdam, The Netherlands: Elsevier, 2011. [Online]. Available: https://books.google.co.uk/books?id=bDtLM8CODsQC
[45] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[46] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[47] R. Duin, "PRTOOLS (Version 3). A MATLAB toolbox for pattern recognition," Pattern Recognit. Group, Delft Univ. Technol., Delft, The Netherlands, 2000.
[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009. doi: 10.1145/1656274.1656278.
[49] M. T. Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality," *Signal Process., Image Commun.*, vol. 52, pp. 20–32, Mar. 2017.
[50] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.
[51] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning: ML Summer Schools*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Germany: Springer, 2004, pp. 63–71. doi: 10.1007/978-3-540-28650-9_4.
[52] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
[53] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1998.
[54] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, "Predicting quality of experience in multimedia streaming," in *Proc. 7th Int. Conf. Adv. Mobile Comput. Multimedia (MoMM)*, Malaysia, 2009, pp. 52–59. doi: 10.1145/1821748.1821766.
[55] A. Bouzerdoum, A. Havstad, and A. Beghdadi, "Image quality assessment using a neural network approach," in *Proc. 4th IEEE Int. Symp. Signal Process. Inf. Technol.*, Rome, Italy, Dec. 2004, pp. 330–333.
[56] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for Internet video," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 339–350, Aug. 2013.
[57] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. Martini, "A classification of video games based on game characteristics linked to video coding complexity," in *Proc. 16th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Amsterdam, The Netherlands, Jun. 2018, pp. 1–6.
[58] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, "NR-GVQM: A no reference gaming video quality metric," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Taichung, Taiwan, Dec. 2018, pp. 131–134.
[59] S. Göring, R. R. R. Rao, and A. Raake, "nofu—A lightweight no-reference pixel based video quality model for gaming content," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6. [Online]. Available: https://stg7.github.io/pdfs/2019/nofu_paper.pdf

NABAJEET BARMAN received the B.Tech. degree in electronics engineering from the National Institute of Technology, Surat, India, in 2011, with a focus on wireless networks, the M.Sc. degree in information technology with specialization in communication engineering and media technology from the Universität Stuttgart, Germany, during which he worked at Bell Labs, Stuttgart, Germany, in 2014, as part of his internship and master's thesis. He is currently pursuing the Ph.D. degree in the field of quality of experience of gaming video streaming applications with the Wireless Multimedia and Networking Research Group (WMN), Kingston University London, where he is currently a Research Associate, involved in QoE-aware video coding strategies as part of MSCA ITN QoE-Net. He is currently a Video Quality Expert Group (VQEG) Board member as a part of Computer Graphics Imagery (CGI) project and is also involved in ITU-T standardization activities. His research interests include wireless networking, multimedia communications, and machine learning.

EMMANUEL JAMMEH received the B.Eng. degree (Hons.) and Ph.D. degrees in electronics systems engineering from the University of Essex, Colchester, U.K, in 1998 and 2005, respectively. He was with the University of Essex, involved in an EPSRC-funded project. He is currently a Research Fellow with Signal Processing and Multimedia Communications Research, Plymouth University, Plymouth, U.K. He worked in various capacities on EU FP7 projects at Plymouth University. He has authored several refereed publications and coauthored the book *Guide to Voice and Video over IP: For Fixed and Mobile Networks* (Springer, 2013). His current research interests include QoS/QoE assessment, control and management, biomedical data analysis and informatics, quality of experience for next generation of emergency communication services, and VoIP quality adaptations.

SEYED ALI GHORASHI received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tehran, Iran, and the Ph.D. degree from King's College London, in 2003. Since 2000, he has been a Research Associate on capacity enhancement methods in multilayer W-CDMA systems sponsored by Mobile VCE, King's College London. In 2006, he joined Samsung Electronics (UK) Ltd and since 2007, he has been serving with the Department of Telecommunications, Faculty of Electrical Engineering, Shahid Beheshti University, G.C., Tehran, Iran. His main area of interest is AI applications in wireless communications.

MARIA G. MARTINI (SM'07) received the Laurea degree *(summa cum laude)* in electronic engineering from the University of Perugia, Italy, in 1998, and the Ph.D. degree in electronics and computer science from the University of Bologna, Italy, in 2002. She is currently a Professor with the Faculty of Science, Engineering, and Computing with Kingston University London, where she also leads the Wireless Multimedia Networking Research Group. She has authored about 150 scientific articles, contributions to standardization groups (IEEE and ITU), and several patents on wireless video. Her research interests include QoE-driven wireless multimedia communications, decision theory, video quality assessment, and medical applications. She has led the KU team in a number of national and international research projects, funded by the European Commission (e.g., OPTIMIX, CONCERTO, QoE-NET, and Qualinet), U.K. research councils, U.K. Technology Strategy Board/InnovateUK, and international industries. She is currently the Associate Editor of the *IEEE Signal Processing Magazine* and was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, from 2014 to 2018. She has also been a Lead Guest Editor for the IEEE JSAC special issue on QoE-aware wireless multimedia systems and a Guest Editor of the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, the *IEEE Multimedia*, and the *International Journal of Telemedicine and Applications*. She chaired/organized a number of conferences and workshops. She is a part of international committees and expert groups, including the NetWorld2020 European Technology Platform Expert Advisory Group, the Video Quality Expert Group (VQEG), and the IEEE Multimedia Communications Technical Committee, where she has served as a Vice-Chair, from 2014 to 2016, as a Chair of the 3D Rendering, Processing, and Communications Interest Group, from 2012 to 2014, and as a Key Member of the QoE and multimedia streaming IG. She is an Expert Evaluator of the European Commission and EPSRC.