Inference Analysis of Video Quality of Experience in Relation with Face Emotion, Video Advertisement, and ITU-T P.1203

Tisa Selma *, Abdelhak Bentaleb *, Mohammad Mehedy Masud *, Saad Harous *

Posted Date: 2 February 2024

doi: 10.20944/preprints202402.0167.v1

Keywords: Quality of Experience, HTTP Adaptive Streaming, Face Emotion Recognition, ITU-T P.1203

Article

Inference Analysis of Video Quality of Experience in Relation with Face Emotion, Video Advertisement, and ITU-T P.1203

Tisa Selma [1,*], Abdelhak Bentaleb [2,*], Mohammad Mehedy Masud [1,*] and Saad Harous [3,*]

[1] Computer Science Department, United Arab Emirates University, Al Ain, UAE
[2] Computer Science and Software Engineering, Concordia University, Canada
[3] Computer Sciences Department, Sharjah University, Sharjah, UAE
* Correspondence: 201990100@uaeu.ac.ae (T.S.); abdelhak.bentaleb@concordia.ca (A.B.); m.masud@uaeu.ac.ae (M.M.M.); harous@sharjah.ac.ae (S.H.)

Abstract: With end-to-end encryption for video streaming services becoming more popular, network administrators face new challenges in preserving network performance and user experience. Video ads may cause traffic congestion and poor Quality of Experience. Because of the natural variation in user interests and network situations, traditional algorithms for increasing QoE may face limitations. To solve this problem, we suggest a novel method that uses user facial emotion recognition to deduce QoE and study the effect of ads. We use open-access Face Emotion Recognition (FER) datasets and extract facial emotion information from actual observers to train machine learning models. Participants were requested to watch ad videos and provide feedback, which will be used for comparison, training, testing, and validation of our suggested technique. Our tests show that our approach beats the ITU-T P.1203 standard in terms of accuracy by 37.1%. Our method provides a hopeful answer to the problem of increasing user engagement and experience in video streaming services.

Keywords: quality of experience; HTTP adaptive streaming; face emotion recognition; ITU-T P.1203

1. Introduction

Video services and related technologies have rapidly advanced at a fast pace in recent decades. Inferring Quality of Experience (QoE) from encrypted data in video streaming apps is a complex task that network service providers must handle. All the abbreviations are listed in Table A1 that can be found in Appendix A. Furthermore, network providers must maintain the highest possible quality by effectively managing traffic and responding to outages in real time. Owing to the limitations of the inference capabilities of encrypted data, traditional techniques that depend on deep packet inspection cannot be used to successfully infer QoE in encrypted network traffic [1]. Consequently, various sophisticated algorithms that use machine learning (ML) methodologies have recently been proposed to forecast QoE indicators [2]. QoE is a metric used to determine how a user experiences a certain service. We believe that good Quality of Service (QoS) can lead to acceptable QoE [3]. However, for deteriorated QoS (high jitter, delay, and packet loss), the chances of obtaining adequate QoE are lower. In this work, objective QoE refers to QoE that can be calculated without any user feedback, and subjective QoE refers to 1–5 user ratings of certain services/multimedia, also known as the Mean Opinion Score (MOS) [4]. Although some psychological behavior, user background, user viewing history, user favorite program and advertisement preferences and other QoE influence factors cannot be controlled, thus, QoS and objective QoE should be prioritized so that we can get optimal final QoE. This inference effort attempts to capture what users like by considering the advertisement position, period duration, and recurring intensity.

 doi:10.20944/preprints202402.0167.v1

The term QoS is used to indicate the level of optimal service provided by some multimedia/service. Currently, the QoE metric indicates the level of user satisfaction experience by considering the user background, psychological–emotional state, and user expectancy level. The QoS usually depends on several objective metrics, including delay variance, jitter, throughput, and delay. The QoE is a subjective matter. This indicates that an optimal QoS metric does not correspond to enhanced QoE. QoE measures the level of user satisfaction with a certain service, which can be determined based on subjective measurements, usually called the mean opinion score (MOS), where a user is asked to rank a video session between 1 (bad) and 5 (excellent).

This study aims to find alternative solutions that can automatically infer end user QoE by observing face behaviors while watching videos to reduce the cost, time, and effort required to perform accurate subjective QoE assessment. Our study proposes an innovative method utilizing facial emotion recognition to infer QoE and examine the impact of ads on user experience. We use our own extracted Face Emotion Recognition (FER) datasets, facial emotion information is extracted from actual observers to train machine learning models. Our approach aligns with psychophysiology based QoE assessment, as explained by Engelke et al. [5], highlighting the importance of understanding the emotional aspects of user experience.

We present some ideas to make automatic video QoE inference more affordable and reliable toward factors that may impact QoE, such as advertisements. We choose advertisement as the impairing factor among the other factors as users encounter advertisements in most video streaming sessions. Furthermore, we aim to investigate how and where advertisements can be shown during video streaming sessions without impairing end-user QoE. We hypothesized that user QoE will be decreasing with an increasing number of advertisements.

We also hypothesize that the better and more stable the video quality, content, and the fewer unrelated advertisements with content, the better the end-user QoE and the more satisfied a user will be. The objectives of our study were as follows:

1. Evaluating the effects of video advertisements on QoE; here, the reliable measurement of QoE is a fundamental step in multimedia communication. Considering all the influencing factors in QoE experiments is important.
2. Compare and analyze the experimental results based on the ITU-T P.1203 standard and existing studies [6].
3. Propose an accurate machine learning model that can estimate QoE by considering advertisements and user expressions.

Acceptable end-user QoE is critical because of video content, and advertisements are widely used. A combination of subjective and objective QoE assessments is a requirement to obtain valid feedback on service performance. The following challenges are found while obtaining credible QoE:

1. Obtaining subjective QoE is expensive in terms of money, effort, and time. Therefore, we attempted to offer an alternative solution to this problem.
2. The ITU-T P.1203 standard QoE measurement did not anticipate some significant factors, leading to inaccurate results. We hope to alleviate this issue using our proposed model.
3. Excessive video advertisements during a video streaming session may weaken user engagement and negatively impact user QoE. We attempted to devise a method wherein a viewing session can coexist with several advertisements if a specific threshold is met.

Based on the above research concerns, the following research questions and their corresponding objectives and contributions can be obtained:

1. How can machine learning be used to infer QoE from different data sources, such as viewers' emotions, bitrates, resolutions, and advertisements?

- We conducted a preliminary experiment and obtained ITU-T P.1203 results with low accuracy.
- We investigated the reasons for the low accuracy of ITU-T P.1203 results.

 doi:10.20944/preprints202402.0167.v1

- We hypothesized that the low accuracy of the ITU-T P.1203 result is because the features considered for building the model are irrelevant to recent conditions, that is, advertisement data and users' facial emotions.
- We proposed video QoE inference by developing an ML model that considers ITU-T standard results, facial expression, gaze behavior, and advertising data.
- We compared various ML models and state-of-the-art algorithms with our proposed model to determine the accuracy of our proposed model.

2. How can high-quality data be collected and prepared to yield an effective ML model for QoE estimations?

- We created and circulated two websites to collect facial and YouTube stats-for-nerds video information.
- We designed more than 100 questions for participants to answer to ensure the accuracy of the obtained data. We asked many questions to make sure that the participants have the same answer about the topic and to minimize the random answer by participant. If we found it random answer or the answer not aligned with the other answer, we simply did not use them.
- We collected 50 video face recordings and metadata from 122 questionnaires from 661 survey participants from 114 countries and cities using an online platform.
- We intended to collect as much complete data as possible from individuals of various ethnicities and nations.

3. How to improve data quality for ML?

- We applied feature extraction and attribute selection to improve the supervised learning methods.
- We preprocessed the data to improve the data quality and thereby improve the ML performance.
- We obtained training data from various sources to enrich the model's accuracy.
- We performed numerous ML experiments and made the necessary modifications to train the model.

4. How can we propose better advertisement duration and placement during video playback to preserve the advertiser budget and boost user QoE?

- We conducted questionnaire research to gather cultural patterns on how, what, and where advertisers can manage their budgets while efficiently promoting products.
- We provided insight into how to compromise video QoE, user engagement, and advertisement efficacy to optimize the budget.

The remainder of this paper is organized as follows. The first section presents the introduction, including the motivation, problem statement, and contribution. The second section presents related work, and the third section presents methodology. The fourth section elaborates on all the gathered data with deeper explanation of the experimental results. Finally, the conclusions are presented in the final section.

2. Background and Related Work

Many OTT services utilize end-to-end encryption for enhanced user privacy and security. However, this encryption utilization may limit the network operator's ability to observe and rectify any quality degradation by employing certain functions such as quality-of-service (QoS) provisioning. Nowadays, several approaches— such as traditional machine learning (ML)–based and session-based models—are used for video QoE inference in encrypted applications [7]. If the QoE and impairments in encrypted networks can be inferred, the impairments can be monitored and addressed. Conventional solutions based on deep packet inspection cannot handle the inference task owing to recently advanced encryption technologies. On-the-fly decryption is unfeasible owing to the rapid development of encryption technologies.

 doi:10.20944/preprints202402.0167.v1

In the rapidly evolving landscape of video streaming services, the increasing popularity of end-to-end encryption presents a set of challenges for network administrators striving to uphold network performance and user experience. The intrusion of video ads into the streaming ecosystem poses a risk of traffic congestion and diminished Quality of Experience (QoE) for users. Traditional algorithms designed to enhance QoE encounter limitations due to the inherent variability in user interests and network conditions. To address this, our study proposes an innovative method employing user facial emotion recognition to deduce QoE and examine the impact of ads on viewer experience. Leveraging open-access Face Emotion Recognition (FER) datasets, facial emotion information is extracted from actual observers to train machine learning models. This approach stands in alignment with psychophysiology based QoE assessment trends, as highlighted by Engelke et al. (2016) [5], acknowledging the importance of understanding the emotional aspects of user experience.

To validate our proposed method, participants are asked to watch an ad video and provide a rating, then use that assessment as a basis for comparison, training, testing, and validation. Our results showed an accuracy improvement of 37.1%, which is much better when compared to the results of ITU-T P.1203. This demonstrates the efficacy of our proposed method in overcoming existing limitations. This is in line with the broader discourse on QoE estimation that incorporates estimation parameters as done by Garcia et al. [6]. They explore the impact of initial loading quality, congestion, and progressive video bitrate.

The research we conducted also contributed to the exploration of dynamic adaptive streaming, as discussed by Pereira and Pereira [7]. They considered the influence of content selection strategies on QoE, as empirically tested by Sackl et al. [10]. In addition, our research is in line with Hoßfeld et al.'s [11] quantitative analysis of YouTube QoE through crowdsourcing. Our research uses adaptive streaming because it significantly improves QoE, as Oyman and Singh explain in their paper in [12]. Meanwhile, Yao et al. in their writing in [13], explained the importance of using real-world bandwidth traces for accurate HAS performance results. In our research, we also used real-world bandwidth traces in line with their testing. We use stats-for-nerds monitoring to evaluate and estimate video QoE experienced by viewers.

In addition, our proposed method is in line with what Ghani and Ajrash did in their [14] article on QoE prediction using alternative methods that do not rely on decrypted traffic data, such as using psychophysiological measures and facial emotion recognition and subjective feedback from viewers in E2E environmental conditions. Meanwhile, Porcu et al. in their writing in [15] conducted research in line with the hypothesis. We believe that QoE end users may be predictable by observing changes in facial emotions, eyes and a multidimensional approach similar to the one we worked on.

While end-to-end encryption protects user privacy in video streaming, it limits network administrators' ability to ensure a positive Quality of Experience (QoE) [12,16]. Traditional methods, relying on network data, go blind in this encrypted landscape, making it unable to identify factors such as intrusive ads that negatively impact service users [7,10]. This is where understanding the user's emotions becomes very important, as expressed by Zinner et al. in their writing in [17].

With many users jockeying for bandwidth, competition to use bandwidth arises and can become into an unfair allocation resulting in QoE not being optimal for all [16]. Our proposed research using FER to predict QoE could potentially reduce this trend by adjusting ad placement and video views based on real-time emotional responses as one of the significant QoE influence factors.

To map audience expressions into accurate QoE measurement requires a robust framework. This is where machine learning algorithms are used. Cohen's Fast Effective Rule Induction (1995) can be used as a powerful tool for extracting meaningful patterns from facial data, allowing us to map emotions to the QoE level [18]. To ensure the reliability of this mapping, agreement between observers needs to be rigorously assessed. The statistics of Kappa Landis and Koch (1977) offer a well-established method for measuring the consistency of human judgments, essential for validating the accuracy of our proposed emotion-to-QoE model [19].

In addition, aligning our QoE assessments with established standards is critical for wider adoption. Bermudez et al. demonstrated the successful application of the ITU-T P.1203 QoE model in

 doi:10.20944/preprints202402.0167.v1

live video streaming over LTE networks [20]. By adapting and integrating this standard framework into our FER-based approach, we recommend compatibility with existing QoE measurement systems, paving the way for seamless integration into video streaming platforms.

The term QoS was previously used to indicate the level of user satisfaction. Nowadays, the QoE metric indicates the level of user satisfaction by considering the user's background, psychological and emotional state, and user expectancy level. The QoS usually depends on several objective metrics, including delay variance or jitter, throughput, and delay. The QoE is a mostly subjective matter, indicating that the optimal QoS metric does not correspond to an enhanced QoE. The QoE measures the level of user satisfaction for a certain service, which can be determined based on a subjective measurement, usually called the Mean Opinion Score (MOS), where a user is asked to rank a video session between 1 (bad) and 5 (excellent). According to a MUX report on video streaming [35], a long rebuffering time is one of the main reasons a user stops watching video content; rebuffering leads to poor image quality and repeated playback errors. Sometimes, users stop watching because of too many advertisements. Many solutions have been proposed to address these problems; however, the level of user engagement or satisfaction remains quite low because these solutions cannot perfectly handle network fluctuations and advertisement problems in a real-time. Herein, we propose breakthroughs in ML that can handle the aforementioned issues by giving predicted insight on what QoE is experienced by users so that video QoS, and advertisement placing scenarios can be optimized to satisfy user QoE and automatically improve overall QoE. The list of content and advert combination can be seen in Table 1 below.

Table 1. Content and advert details.

Content Title	Content Length (s)	Number of Ad	Length of Ad	Position of Ad
Expo 2020 Dubai	280	1	18	Post-roll
Squid game2	113	1	30	Pre-roll
Every death game SG	375	1	18	Mid-roll
5 metaverse	461	3	75	Pre-roll, mid-roll, post-roll
Created Light from Trash	297	2	45	Pre-roll
How this guy found a stolen car!	171	6	288	Pre-roll, mid-roll, post-roll
First underwater farm	233	6	198	Pre-roll, mid-roll, post-roll
Most beautiful building in the world	166	6	292	Mid-roll
This is made of...pee?!	78	4	418	Pre-roll
The most unexplored place in the world	256	5	391	Post-roll
Jeda Rodja 1	387	8	279	Pre-roll
Jeda Rodja 2	320	8	440	Pre-roll, mid-roll, post-roll
Jeda Rodja 3	415	6	272	Pre-roll, mid-roll, post-roll
Jeda Rodja 4	371	6	311	Post-roll
Jeda Rodja 5	376	6	311	Mid-roll

2.1. Quality of Experience

According to the ITU-T standard, QoE is "the overall acceptability of an application or service, as perceived subjectively by the end-user" [25]. It is inherently subjective, as it is based on a user's perspective and the user's own idea of "high quality." The ability to assess QoE will provide network operators with a sense of the network's contribution to total customer satisfaction in terms of

 doi:10.20944/preprints202402.0167.v1

dependability, availability, scalability, speed, accuracy, and efficiency. Thus, many network researchers are now working on this topic and are attempting to incorporate it into network choices to ensure high customer satisfaction while using minimal network resources.

Mean opining score (MOS) is an example of a subjective measuring approach, wherein consumers rate service quality by assigning five distinct point ratings ranging from 1 to 5, with 5 being the best and 1 being the worst. MOS represents discrete values; however, it can be expanded to non-discrete values. The opinion score (OS) has been proposed as a new measure of QoE, with a new value of 0. According to the OS scale, quality is defined as awful (0–1), poor (1–2), fair (2–3), good (3–4), or exceptional (4–5). First, defining the components that affect QoE is vital. QoS primarily affects user experience (UE) because numerous QoS factors directly or indirectly affect user perceived QoS. The key QoS parameters that impact multimedia services are the bandwidth, jitter, delay, and packet loss rate.

Furthermore, in Table 2, we can see the foundation of our proposal is fortified by insights from many studies in related areas. Amour et al. [23] introduce an improved QoE estimation method based on QoS and affective computing, emphasizing the relevance of emotional factors in user experience. Bhattacharya et al. [24] highlighting an affect-based approach in the evaluation of audio communication QoE, further strengthening our focus on emotions in assessing end user experience. Porcu et al. [15,26] and Antons et al. [27] provide notable point of view into estimating QoE using facial emotion gesture and electroencephalography, respectively, reinforcing the importance of multimodal approaches.

Moreover, the utilization of EEG correlates during video quality perception as mentioned by Kroupi et al. [28]. Moreover, eye-tracking combined with correlates of brain activity to predict quality scores as elaborated by Arndt et al. in [29,30] underscores the interdisciplinary nature of our proposed method. Additionally, research on the role of spatio-temporal distortions as explained by Engelke et al. in their work in [31] and gaze disruptions in nonuniformly coded natural scenes [32,33] in line with our aim to understanding the impact of video content on user attention.

Furthermore, the real-time classification of evoked facial emotions using significant facial feature tracking and physiological responses [34] contributes to the broader context of emotion-aware computing, supporting our approach to leveraging emotional cues for QoE prediction.

Table 2. Our contribution among other related works.

Reference	Influence Factors	Considered Features
Amour et al. [23]	Resolution, bandwidth, delay	Face emotion
Bhattacharya et al. [24]	Delay, packet loss, bandwidth	Acoustic feature
Porcu et al. [15]	Delay, stalling	Face and gaze tracking information
Porcu et al. [26]	Blurring	Face and gaze tracking information
Antons et al. [27]	Noise signal	EEG
Arndt et al. [29]	Low bitrate encoding	EEG and EOG
Kroupi et al. [28]	High quantization attribute	EEG
Arndt et al. [30]	Low bitrate encoding	EEG and gaze movement information
Rai et al. [32]	High quantization attribute	Gaze movement information
Rai et al. [33]	Packet loss	Gaze movement information
Engelke et al. [31]	Packet loss	Gaze movement information
Bailenson et al. [34]	Provoked delight and anxiety	Face emotion and 15 physiological features
Our Proposed Work	Video advertisement	Face emotion, video metadata and advertisement information

 doi:10.20944/preprints202402.0167.v1

2.2. ITU-T P.1203 Standard

We compare our results with those of the ITU-T P. 1203 standard algorithm and several state-of-the-art machine learning algorithms. ITU-T P.1203 is the first standardized QoE model for audiovisual HAS and has been thoroughly trained and verified on over a thousand audiovisual sequences with HAS-typical effects, such as stalling, coding artifacts, and quality switches. At the bitstream feature level, the ITU-T P. 1203 dataset contained four of the 30 approved subjective databases. For video quality analysis, it uses bitstream-based models over metadata-based models and mixes classical models with machine learning-based techniques to predict user QoE. [35]

The P.1203 set of standards [36] developed by ITU-T is an example of a bitstream-based model. P.1203 is a quality model for HTTP-based adaptive audiovisual streaming [36–38]. It is divided into three sections: Pv, short-term video quality prediction; Pa, audio short-term quality prediction; and Pq, total integration of quality, including the perceived stalling effects. The Pv module in P.1203 has four different operation modes, ranging from mode 0 to mode 3. The modes are defined by the amount of bitstream information provided, ranging only from the metadata (codec, resolution, bitrate frame rate, and segment time) in mode 0 to complete bitstream access in mode 3.

Herein, we focus only on the mode 0 Pv model. Advantageously, mode 0 only requires metadata and is the quickest of all modes. However, the accuracy of Mode 0 was lower than that of Mode 3. In contrast, Mode 3 requires a patched client decoder that extracts bitstream properties, such as QP values. The present P.1203 standard does not consider newer codecs such as H.265, VP9, and AV13, which are being utilized in DASH streaming. Furthermore, it is limited to resolutions of up to 1080 p and frame rates of up to 24 fps.

In addition, the standard has low accuracy in face emotion recognition (FER) and MOS. Hence, this study aims to improve the accuracy by considering the ITU-T P.1203 standard's results, the advertisement effect as another QoE influence factor (IF), and the FER results.

2.3. Face Emotion Recognition (FER)

One of the most important areas of research on human–computer interaction and human emotion detection is facial expression recognition [39]. A system must process various variations of the human face to detect facial expressions such as color, texture, posture, expression, and orientation. To determine a person's facial expressions, various facial movements of the muscles beneath the eyes, nose, and lips were first detected and then categorized by comparison with a set of training data values using a classifier for emotion recognition. We used face behaviors along with advertisement insertion information as significant IFs, in addition to other QoE IFs, to enable automatic QoE assessment. In the future, we intend to remove the subjective QoE assessment that requires the user to fill out several questionnaires and provide ratings on a 0–5-star scale.

2.4. HTTP Adaptive Streaming (HAS)

YouTube is one of the most popular examples of a HAS application. YouTube has always used server-based streaming, but it has recently added HAS [40] as its default delivery/playout technique. HAS requires a video to be accessible at numerous bit rates, that is, different quality levels/representations, and to be divided into short chunks of a few seconds each. The client evaluates the current bandwidth and/or buffer state, requests the next segment of the video at an appropriate bit rate to avoid stalling (i.e., playback stoppage due to empty playout buffers), and best utilizes the available bandwidth.

HAS is based on traditional HTTP video streaming and allows changes in video quality during playback to adapt to changing network circumstances. On the server, the video was divided into separate segments, each of which was available in different quality levels/representations (representing different bit rate levels). Based on network measurements, the client-side adaptation algorithm requests the next segment of the video at a bit rate level appropriate for the current network conditions [41].

 doi:10.20944/preprints202402.0167.v1

2.5. QoE Influence Factors

QoE is an important factor for determining user satisfaction with advertisements. This is an important indicator of the psychological expectations of a user's fulfillment. To meet the user expectations for a high QoE, we can elaborate on several metrics [42,43], such as human IFs, which have the most complex parameters, like system IFs. Human IFs include all user information. Contextual IFs include information about location, user premise (watching environment), time of watching (day or night), type of usage (casual watching, newly released favorite online gaming video, etc.), and consumption time (offload time and peak hours). System IFs include technical factors of video quality that can be quantitatively measured using QoS measurements, such as delay, jitter, packet loss, and throughput. Content IFs comprise characteristics and information about the content of the video being watched by users.

2.6. QoE Metrics

Several standards are used by ITU-T and ITU-R for video QoE subjective test settings, audio-visual video streaming, and subjective QoE grading techniques. Accordingly, some grading techniques and testing standards have been proposed to enhance viewer satisfaction levels and MOS. Several factors may affect the QoE assessment.

- Source video quality: Content quality may be affected by the characteristics of the original video, such as codec type and video bitrate.
- QoS: mainly the consideration of how the packet or video traffic chunks travel in the network from the source to the destination; alternately, technical details include packet loss, jitter, delay, and throughput.
- MOS or subjective QoE measurement: this involves the human perception or satisfaction level.
- Objective QoE measurement: This denotes assessment models for estimating/predicting subjective video quality services by extracting important QoE metrics, for example, by examining the stalling frequency and stalling period.

2.7. QoE Assessment Types

As mentioned previously, QoE assessment techniques can be divided into subjective, objective, and hybrid assessments. Subjective measurements are time-consuming but can directly measure the user satisfaction level. They usually consider the user's preferences, age, individual psychology, viewing history, and so on. Additionally, they fully depend on each user's perspective and differ depending on the user; hence, they are processed using mathematical models (i.e., mean, standard deviation, and regression) to handle per-perspective bias. Subjective measurements can be categorized into single and double stimuli, based on the presence of samples.

In the objective QoE measurements, real-time quality assessments are performed using a computational model to estimate the subjective test results. They aimed to predict an MOS that is as similar as possible to the real MOS obtained via subjective QoE measurements. The root-mean-square error (RMSE) [21] and Pearson correlation are common metrics used to determine the relationship between objective (predicted MOS) and subjective (real subjective MOS) measurements. Objective measurements were divided into audio and video quality measurements [22]. Audio quality measurements include parametric, non-intrusive, and intrusive techniques. In intrusive techniques, the original signal is compared with the signal degraded in a test case. This technique yields accurate results but cannot be performed in real-time. The non-intrusive technique estimates the audio quality using only the degraded signal. The parametric technique can predict audio quality using network design process characteristics/attributes such as echo, loudness, and packet loss [44].

Hybrid assessments have the advantages of both subjective and objective practicality, portability, and convenience. Garcia et al. [45] defined an optimal approach for QoE assessment of multiple-description coding in video streaming in an overlay network. They discussed the hybrid pseudo-subjectivity of the video quality assessment approach, which outperformed the peak signal-to-noise ratio (PSNR) based on experiments. Subjective assessment is the most precise assessment

 doi:10.20944/preprints202402.0167.v1

method because it can accurately and directly collect information about user perceptions, personal expectations, and UE. Although the subjective assessment method has many advantages, it also has some limitations. For example, many factors must be considered, complex procedures must be implemented, considerable data and human resources are required, and the method cannot be applied to real-time applications. Owing to these limitations, the subjective assessment method is not widely used, except for verification purposes and comparisons with other methods.

However, an objective assessment method is both convenient and practical. It can be formulated as a mathematical model considering various QoS metrics, including jitter, packet loss, throughput, and delay. However, it yields low accuracy because it denotes the approximate quality experienced by the user, but not the user's real experience and expectations.

By considering user behavior when watching a video, the hybrid assessment method can be implemented in real-time with high accuracy owing to supporting factors such as artificial intelligence and statistics that can reduce the limitations of both subjective and objective assessment methods. However, it requires a considerable amount of data and complex model training, computation, and validation. ITU-T has published the ITU-T Rec. P.1203, wherein an objective model was proposed for measuring QoE in video-streaming systems. To comply with the ITU standard, we used MOS as a standard comparison with the five-scale absolute category rating (ACR), as shown in Table 3. It is an ITU 5-point quality scale. [35].

Table 3. Absolute Category Ranking (ACR) score.

Grading Value	Emotion
5	Happy
4	Surprised
3	Neutral
2	Sad, fear
1	disgust, anger

3. Methodology

The overall proposed work is shown in the Figure 1 and consists of five steps. The steps are as follows, first, video watching session, second, data collection and storing, third, combining all data approaches, fourth, data preprocessing, data cleaning, training, and model evaluation, and the last step is QoE estimation results and analysis.

These steps will be elaborated in detail later in Sections 3.1 until 3.5. Meanwhile, to answer research questions number two (How to collect and prepare high-quality data to yield an effective ML model for QoE estimation?) mentioned in the introduction, we collected data from many sources to obtain many possibilities and various answers to better train our model. Seven hundred participants from 114 different countries completed the questionnaires, of which 125 were valid. Face and screen recordings were obtained for 60 participants, and 40 pairs of face and screen recordings were obtained.

 doi:10.20944/preprints202402.0167.v1

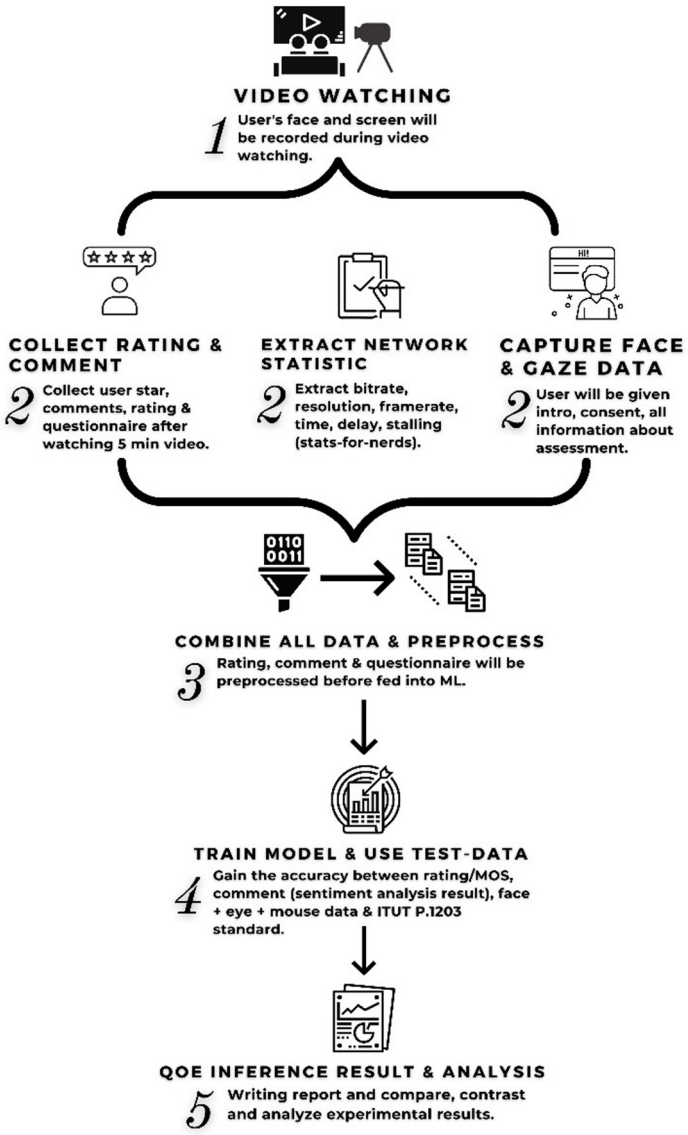


Figure 1. Machine learning process summary.

Some of the data, both video recording and questionnaire were eliminated due to data cleaning, pre-processing, and separating with invalid data, blur/low brightness videos, and incompatible results. The methodology used in this study is illustrated in Figure 2. During the initial stage, as shown in Figure 2, the user enters the laboratory room and reads the agreement. If they accept it, they proceed with it. If they do not accept it, they can leave at any time and for any reason. We recruited participants by advertising the questionnaire and survey to be taken by individuals wanting to earn some money (around 0.5 USD).

 doi:10.20944/preprints202402.0167.v1

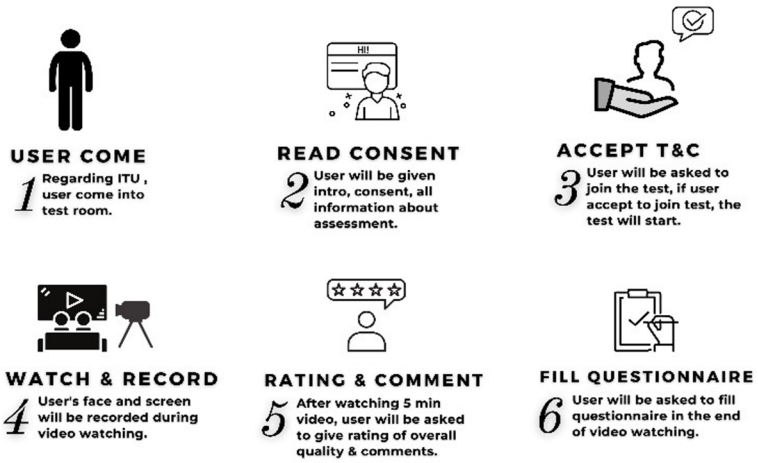


Figure 2. Data collection process.

All platforms are centralized into one www.tisaselma.online/form.html for Indonesians who talk in Bahasa Indonesia or www.tisaselma.online/form2.html for English users. Moreover, the automatic platform www.xavyr.co.uk (closed for public use) was developed for automatic face and screen recordings. The details of the contents and advertisements are summarized in Table 1.

Moreover, if the participants gave their consent, we videotaped their face and screen during the video-viewing session. After watching the video, the participants were prompted to rate it from 1 to 5 and leave a remark. They were asked to complete a questionnaire at the end of the 30 min video viewing session. The recording was then pre-processed and fed into the ML and DeepFace algorithms. Furthermore, we extracted the stats-for-nerds, which were fed into the ITU-T P.1203 model.

All the questionnaire questions were provided in English and Bahasa to enable broader participation. The English and Bahasa forms were identical. Only the video content titles and advertisements integrated into them were different. Details of the video content titles, and advertisement data are provided in Section 4, Table 9. We choose the advertisement as a disturbance factor that do not have any relation with the content, have many repetitions, located in the beginning, middle or end of content and they are free ads that can be downloaded in internet for free. All participants were asked to watch five videos for approximately five minutes without advertisements. The advertisements were shown for approximately 7–11 min.

3.1. Video Watching Session

If participants agreed about the terms and conditions, we let them watch the video while collecting video recordings of the face and screen. The process is summarized in Figure 2.

3.2. Data Collection and Storing

All star ratings and comments were collected from what they watched after watching each video. Some of the questions can be found in Appendix B. The complete results and their analysis we plan to be elaborated in another journal.

3.3. Combine All Data Approach

From the video screen recording, we can extract Stats-for-Nerds to obtain the bit rate, resolution, and frame rate for every 4 s window. A 4 s window was utilized as a period to ease the prediction process by down sampling. To extract the video statistics, we capture the stats for nerds that contain bitrate, buffer health condition, bandwidth, frame per rate, resolution, and video ID. To extract stats for nerds that will be the input for the ITU-T P.1203 model, we can perform a right click on YouTube player, and choose the stats-for-nerds option to be turned on accordingly, as shown in Figure 3, the stats-for-nerds statistics shown as seen in Figure. 4 and the result of stats-for-nerds extraction input

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2024 doi:10.20944/preprints202402.0167.v1

and output are shown in Figure 5. This information can be observed if we click on the YouTube video player during video playback. From time to time, we captured every second of network condition statistics during the video watching period playback.

When the user is watching, we capture the video statistics and at the same time we record the user face video. We used VLC to extract the frame from the video. Then, we obtain the information that needs to be fed into the ML model, that is, bitrate, frame rate, and resolution. The resulting features extracted from stats-for-nerds for each video from all participants are listed in Table 4. Next, we feed the video statistics, including all video metadata (*. json file) in ITU-T 1203 machine learning to predict the Mean Opinion Score (MOS). The ITU-T 1203 machine learning model no longer needs to be trained during this step.

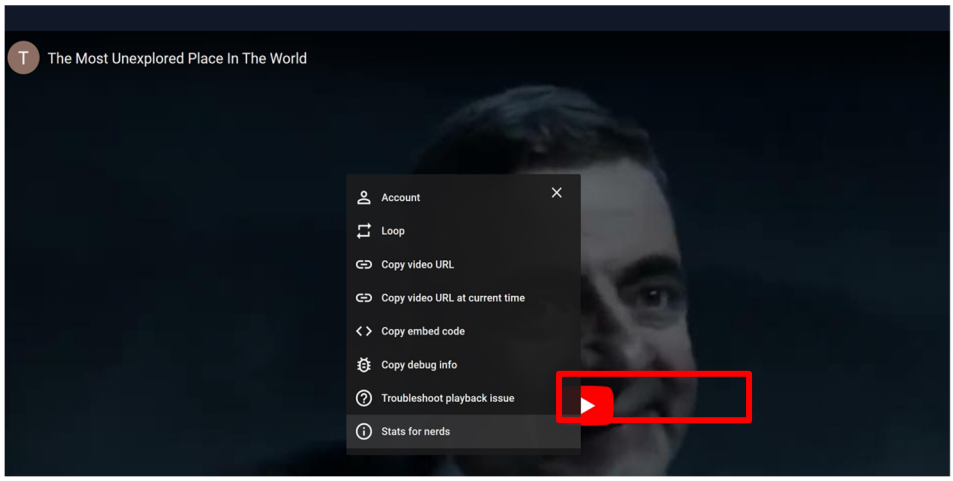


Figure 3. Stats-for-Nerds on YouTube player by right clicking the player.

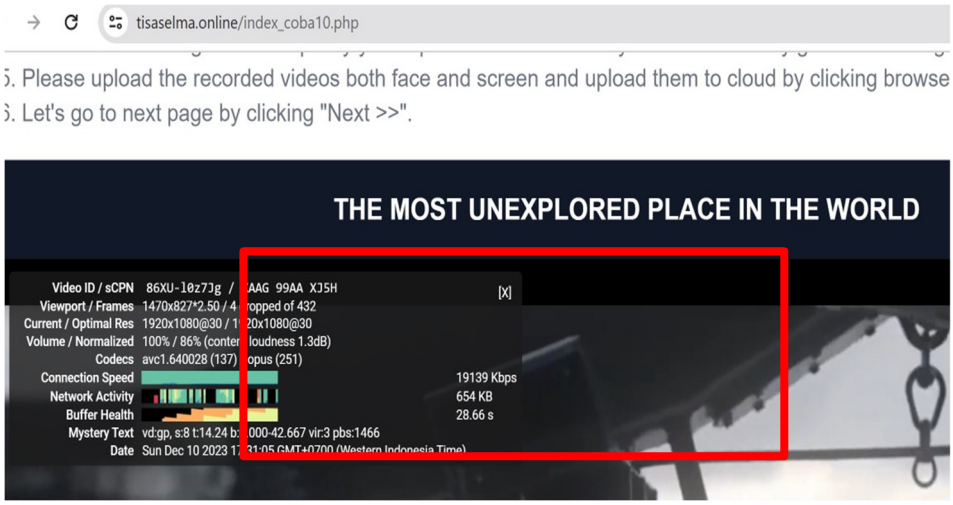
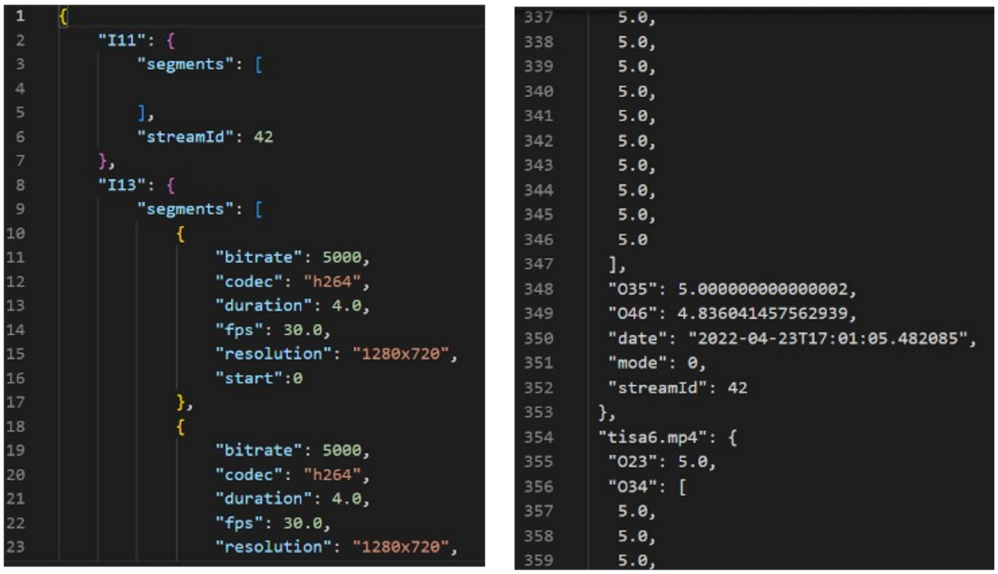


Figure 4. Stats-for-Nerds appears in the right above corner of YouTube player.

 doi:10.20944/preprints202402.0167.v1



```
{
    "I11": {
        "segments": [

        ],
        "streamId": 42
    },
    "I13": {
        "segments": [
            {
                "bitrate": 5000,
                "codec": "h264",
                "duration": 4.0,
                "fps": 30.0,
                "resolution": "1280x720",
                "start":0
            },
            {
                "bitrate": 5000,
                "codec": "h264",
                "duration": 4.0,
                "fps": 30.0,
                "resolution": "1280x720",
```

```
    5.0,
    5.0,
    5.0,
    5.0,
    5.0,
    5.0,
    5.0,
    5.0,
    5.0,
    5.0
   ],
   "O35": 5.000000000000002,
   "O46": 4.836041457562939,
   "date": "2022-04-23T17:01:05.482085",
   "mode": 0,
   "streamId": 42
  },
  "tisa6.mp4": {
   "O23": 5.0,
   "O34": [
    5.0,
    5.0,
    5.0,
```


Figure 5. The screenshot of input and output *.json file from ITU-T P.1203. in the left is the input and the right is the output file.

Table 4. Sample features extracted from stats-for-nerds for each video from all participants.

Resolution	Bitrate	ITU-T P.1203 Results	Video Content Length	Star Review
1080	8000	5	301	1
1080	8000	5	301	2
1080	8000	5	301	1
1080	8000	5	301	4
1080	8000	5	301	4
720	5000	5	301	5
720	5000	5	303	1

Our main competing approach was the ITU P.1203 standard. The input features for the ITU were *. json file that contains a frame rate, bit rate, resolution, stalling length, and stalling position. From the input file in Figure 5, we can see that for every 4 s window there may be fluctuating frame rate, bit rate, resolutions, stalling length, and position. This input of every 4 seconds window is obtained from YouTube stats-for-nerds every 4 seconds' window. This input from the 4 s window will be given 1–5 ACR score from ITU-T P.1203 model. The following standards were implemented using assessment software [3,45]:

1. P.1203 (ITU-T): Parametric bitstream-based quality evaluation of progressive download and adaptive audiovisual streaming services over dependable transport.
2. P.1203.1, ITU-T: Parametric bitstream-based quality evaluation of progressive download and adaptive audio-visual streaming services over dependable transport–video quality estimate module.
3. ITU-T Rec. P.1203.2: Audio quality estimate module for metric bitstream-based quality evaluation of progressive download and adaptive audiovisual streaming services over dependable transport.
4. ITU-T Rec. P.1203.3: Quality integration module for metric bitstream-based quality evaluation of progressive download and adaptive audiovisual streaming services over dependable transport.

 doi:10.20944/preprints202402.0167.v1

3.4. Pre-processing, Data Cleaning, Model Training and Evaluation

We performed data cleaning, data preprocessing, and feature selection to obtain better data quality and remove all unreadable data from our dataset. First, we manually checked all the video recordings; if there was insufficient lighting or the participants used hats and sunglasses during video recording, the machine learning model would not detect the emotion properly. Hence, we remove this information from our dataset.

For the preprocessing step, we combined all the features from the video statistics, star rating, and emotion recognition for comparison using the same ACR score. After extracting frames from the face video recording using VLC, we put all the resulting frames into DeepFace machine learning and performed several training and testing steps. For attribute selection, we selected the eight best attributes using symmetrical t attribute val, Ranker using 10-fold cross validation, we take the best nine Selected Attributes using Relief F Attribute Eval, Ranker using 10-fold cross validation, and we take 13 selected attributes using correlation attribute val using Ranker using 10-fold cross validation. These attributes are: long 5min ad, FER, ad loc, name, bitrate, length ad, resolution, title, ad count, content length, ITU-res, ad each min, repeat.

The frame is extracted from the face videos to be fed into the CNN/DeepFace model for FER. In this step, we trained the model to satisfy certain conditions. The output includes seven emotions: happy, surprised, disgusted, sad, fear, neutral, and anger. We mapped all emotions to MOS regarding their relatedness. The ACR score can be seen in Table 2 and the mapping from DeepFace emotion to ACR score can be seen in Table 5.

Table 5. Estimated emotion mapping to ACR.

Grade	Estimated Quality	Estimated Emotion
5	Excellent	Happy
4	Good	Surprise
3	Fair	Neutral
2	Poor	Sad
1	Bad	Disgust, Anger, Fear

On the other hand, due to limitation in time and cost, we will try to improve the dataset in our future works. It is quite difficult to get large participants who give consent to give their face recording to be used in our research. It takes around half a year to one year to gather more than 600 participants to fill our questionnaire and 50 participants to give their face and screen recording. Some of the participants come to our laboratory, and some other participants were conducting the test online using unified our self-made platform.

Nevertheless, collecting the required data poses a considerable challenge. Our dataset possesses a unique nature, and we have not identified any existing dataset that aligns directly with the specifics of our research. Considering the constraints of time and budget, augmenting our dataset remains a focal point for our future initiatives. The process of securing consent from a significant number of participants for face recordings is demanding, typically requiring six months to a year to gather over 600 questionnaire responses and obtain face recordings from 50 participants. By 50 participants, we mean, they watched 5 videos of 10 minutes. The results we have 50 multiplied by 5 videos equal 250 videos. Each video contains almost 10 minutes before cleaning and preprocessing. The raw video may contain almost two hours of video face recording sessions. It excludes the screen recording sessions while they are watching video. We have around 500 videos in total. The screen recording video is obtained to know where their attention and perspective is while they are using our platform. We obtained this data for user engagement optimization in the future.

In the future, to get better accuracy and a more robust face emotion recognition model, we will try to expand and vary our video dataset. Firstly, we have generated 1 million synthetic data generated with generative AI model, but we need more time to figure the best hyperparameter to tune with our model. Hence, we plan to publish the result in our future work. Next, we plan to do emotion simulation by utilizing tools and techniques like facial action coding to generate synthetic

 doi:10.20944/preprints202402.0167.v1

facial expressions that represent diverse emotions and data augmentation to artificially increase the size and diversity of our dataset without collecting new data. Additionally, we will organize our data to implement a well-structured data organization system for efficient annotation, analysis, and model training. Moreover, we will balance our data by applying a cost sensitive support vector machine (CS-SVM).

3.5. QoE Evaluation Result and Analysis

The six emotions from DeepFace are mapped to the ACR score, as shown in Table 5. From DeepFace, we obtained a prediction of MOS from 1-5. These networks extract features from questionnaires, such as demographics and user preferences for ads, videos, placement, ad match to video content, and so on. User preferences can reveal users' expectations and perceptions of the expected quality of streaming video services. The output from these three networks is then fed into the QoE prediction network. This QoE prediction network is used to predict QoE scores in relation to advertisements, video content, network conditions and users' emotional states. QoE prediction networks are trained using QoE scores collected from star ratings that users provide after watching each video. We evaluated our proposed architecture using user-generated face recording, video ads, questionnaires, and QoE star rating scores. We compare our architecture with traditional QoE prediction methods and show that our architecture achieves much better results. In addition, our architecture can capture the complex relationship between QoE and the factors that influence it, such as facial expressions, video ads content, and user preferences. This ability to capture complex relationships is due to the use of deep learning and machine learning algorithms, which can learn non-linear relationships from raw data.

On the other hand, as mentioned in the architecture, we use DeepFace to predict the user emotion automatically. DeepFace is a deep learning-based facial recognition system developed in 2014 by Facebook Research that has high accuracy up to 97.35% on the Labeled Faces in the Wild (LFW), which is better than human performance. The DeepFace architecture consists of several convolutional and pooling layers, followed by three locally connected layers and a final softmax layer. The input to the network is a 152x152 RGB image of a face. Our proposed architecture can be seen in Figure 6 below.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2024 doi:10.20944/preprints202402.0167.v1

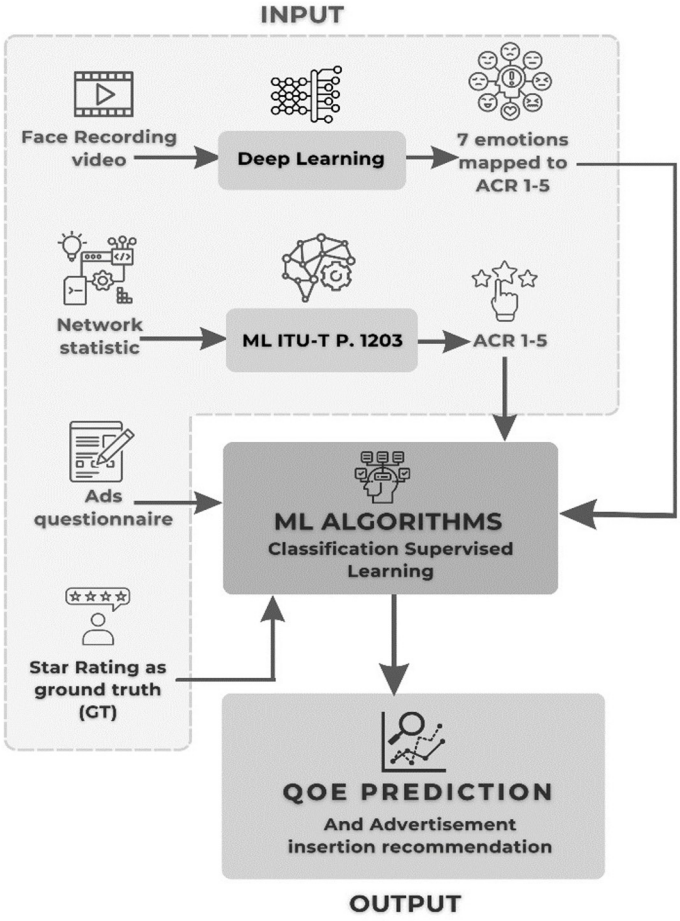


Figure 6. Model Architecture.

Firstly, the network extracts low-level features from the image using convolutional layers. Secondly, these features are then down sampled using pooling layers. The locally connected layers are used to extract higher-level features from the low-level features. Thirdly, the softmax layer is used to classify the input image as one of the faces/emotions in the training dataset. DeepFace was trained using stochastic gradient descent on a dataset of over 4 million faces. The dataset includes faces ranging from many ages, ethnicities, and genders.

Moreover, DeepFace is implemented using Caffe deep learning framework and use Rectified linear units (ReLUs) as activation functions. It uses pooling layers with max pooling with a stride of 2 with locally connected layers use a 3x3 kernel size. The softmax layer has 128 outputs, related to the number of faces in the training dataset.

3.7. Analysis on Machine Learning Methodologies, Features Importance, and QoE Perceptions

In the context of improving end user video Quality of Experience (QoE) perception, leveraging machine learning explanations and feature importance analysis can indeed provide valuable insights. Here's a more in-depth analysis of how these aspects could achieve better QoE perception using face emotion recognition, advertisement insertion information, and network conditions.

In face emotion recognition, feature importance in emotion recognition is to identify key facial features consisting how to understand which facial features have the highest impact to emotion recognition that can enhance the interpretability of our model. By building our proposed model, it can extract how facial expression can affect the perceived content quality. Our future work will work on real time adjustment by developing a system that tracks dynamic changes in facial expression. It will enable real-time adjustments to content delivery, such as aligning the content to match the end user's emotional state.

 doi:10.20944/preprints202402.0167.v1

Our proposed model leveraging multimodal analysis by integrating face emotion recognition, ad insertion details, and network conditions into a holistic QoE model to get a more comprehensive view of the end user experience. uses identification on relevant features by analyzing which features contribute most to determining ad relevance, such as context, viewer demographics, or emotional state that can help optimize ad insertion strategies. By analyzing the questionnaire results, we found that most users hate the un-skippable ads about 32%. With this questionnaire results, we propose a personalized ad delivery by implementing adaptive content insertion that can adapt ad content based on historical user preferences, emotions, and contextual information, resulting to a more personalized and engaging experience. Moreover, by considering network conditions while watching sessions, we can infer the user perception on QoE. We investigate the relation between content delivery strategies and ad insertion scenario in relation to user star rating. We found that the user can still be happy while watching ads up to 10 seconds about 43%. We elaborate this questionnaire results in another journal.

We use neutral content and ads to clarify the face emotion triggered either by network conditions or advertisement scenarios. Emotion related to network conditions may be triggered negatively if the network (bandwidth, resolution, latency, stalling time, stalling frequency, etc.) gets worse. And we can see emotion related to advertisement insertion may be influenced by many advertisements in mid-roll, un-skippable ads, or too long unrelated ads during watching session.

Moreover, feature importance and accuracy are two crucial aspects in the field of machine learning and data analysis. The relationship between feature importance and accuracy has been a subject of interest in various domains, including medicine, computer science, and artificial intelligence. Several studies have explored this relationship and have provided valuable insights. Han & Yu in [46] provided a theoretical framework explaining the relationship between the stability and accuracy of feature selection, emphasizing the dependency of feature selection stability on sample size. This highlights the importance of considering the stability of feature selection methods in relation to accuracy, especially in the context of varying sample sizes.

Strobl et al. in their work in [47] highlighted a bias in variable importance measures towards correlated predictor variables. This bias can impact the accuracy of feature selection methods, indicating the need to account for correlations among predictor variables when assessing feature importance to ensure accurate and reliable results. Furthermore, Altmann et al. in [48] introduced a corrected feature importance measure, emphasizing the importance of using accurate and reliable feature importance measures to ensure the effectiveness of feature selection methods in improving accuracy.

Moreover, Menze et al in [49] compared random forest and its Gini importance with standard chemometric methods for feature selection and classification of spectral data, indicating the preference for Gini feature importance as a ranking criterion due to its consideration of conditional higher-order interactions between variables, which can lead to better accuracy in feature selection. Overall, the relationship between feature importance and accuracy is multifaceted, involving considerations such as stability, bias, correlation among predictor variables, and the impact of feature selection on classification accuracy. Understanding and addressing these factors are essential for optimizing feature selection methods and improving the accuracy of machine learning models.

4. Experimental Results

4.1. Survey Results and Statistics

For the advertisement investigation, we designed questions for participants to ensure the accuracy of the obtained data. The results of this survey are summarized in Tables 6–8.

 doi:10.20944/preprints202402.0167.v1

Table 6. Questionnaire results: the most annoying advertisement type.

The Most Annoying Advertisement Type From 122 Participants		
Case	**Participants**	**Percentage (%)**
Many repeated advertisements in 1 point of time in mid-roll	22	18.03%
Single 5 minutes advertisement long in mid-roll	22	18.03%
In 5 minutes of video content, every 1 minute there is one repeated ad.	21	17.21%
The same advertisement is repeated in pre, mid, and post-roll.	18	14.75%
There is no skippable advertisement.	39	31.97%
Total	**122**	**100%**

Table 7. Survey summary of joining participants.

Amount	Types
661	Total participants from around the world
114	Countries and cities
125	Completed questionnaires
30	Questionnaires completed with video recordings

Table 8. Questionnaire results: the maximum acceptable ad length.

The Maximum Acceptable Advert Length Period	
Time	Participants
< 10 second	41.67%
10 – 30 second	37.5%

In Table 9, we can see that title, video resolution, video bitrate, ITU-T P.1203 result (ACR score), face emotion recognition result (ACR score), content length (in second), the number of advertisements happened in one video content, the length of advertisement (in second), advertisement location, repeated advertisement (1 is for repeating and 0 is no repetition of advertisement), the presence of five minutes length advertisement in one video content (1 is for present, 0 is for absent), the presence of advertisement in each minutes of video content (1 is for present, 0 is for absent), the same advertisement used in pre-roll mid-roll and post-roll (1 is for present, 0 is for absent), the presence of un-skippable advertisement in one video content (1 is for present, 0 is for absent), and the last is the test case or ground truth in supervised classification: star rating 1-5. All these parameters are used in machine learning to predict end user QoE.

Table 9. Sample list of features including ITU-T P. 1203, video metadata, FER values and star rating. We write here only partly because of page limitations. As for the complete data, it is up to hundreds of rows. For a condensed version of the comparison between ITU-T P.1203 results, face emotion recognition result, star rating as ground truth given by participant and our predictions can be seen in Table C1 in Appendix C.

Title	Res	Bitrate	ITU	FER	Cont. length	Ad. count	Long. ad	Ad. loc	repeat	5min. len.ad	Ad.each .min	p/m/p same ad	No skip ad	Star
Stolen_car	720	5000	5	3	459	6	288	4	0	0	1	0	1	1
Underwater_farm	720	5000	5	3	431	6	198	4	1	0	0	1	1	2
beautiful_building	720	5000	5	3	608	6	442	2	0	0	0	0	1	1
Made_of_pee	720	5000	5	3	496	4	418	1	0	0	0	0	1	4
Unexplored_place	720	5000	5	3	433	4	177	3	0	0	0	0	1	4

 doi:10.20944/preprints202402.0167.v1

a. Competing Approaches

We used several machine-learning models to investigate their accuracy against our dataset, including TreeSPAARC. Random Forest, Tree Optimized Forest, Local KNN, Multi-Layer Perceptron, Naive Bayes Simple, Meta-Ensemble Collection, Rules JRip, Rules Furia, Naive Bayes Updatable, Multi-Layer Perceptron CS, Meta-Random Subspace, Chirp, Multiclass Classifier, Meta-Decorate, and SMO.

4.3. Hardware and Software Setup

In brief, we performed almost all experiments using Weka version 3.8.6 in Windows 11 Pro, 21H2, Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz. Some notable parameters used in our models are face emotion recognition (1-5 value), star rating (as a target class 1-5 value), ITU (1-5 value), video statistics (bitrate, resolution, stalling, delay, etc.), and advertisement metadata and information.

4.4. Evaluation

In this section, we elaborate more detail about the experimental results in detail. Our evaluation method used 80:20, 94:6, and 92:8 Pareto for training and testing split dataset, cross-validation 50 and 60 folds. We used this split evaluation method after several attempts to obtain the highest possible results, and the results shown in the table are the highest among all trial error experiments. To evaluate the accuracy of our approach, we used correctly classified instances and incorrectly classified instances, Root Mean Squared Error (RMSE), Precision and Recall. The evaluation results are summarized in Table 10.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

The better model has the least error, similar to the RMSE; the better the model, the lower the error or the lower the RMSE value [33]. Correctly classified instances graph can be seen in Figure 7a. Summary of RMSE results can be seen in Figure 7b, summary of precision and recall can be seen in Figure 7c and Figure 7d Precision is defined as the ratio of True Positives to all Positives. In our issue statement, it would be the proportion of star ratings accurately identified as subjective star rating MOS among all participants who really had it. Mathematically, precision can be formulated as true positives divided by the summation of true positives and false positives.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

On the other hand, recall is a measure of how well our model identifies True Positives. Thus, recall informs us of how many star ratings we accurately recognized as real star rating values out of all those participants. Mathematically, recall is the true positive divided by the summation of true positives and false negatives.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

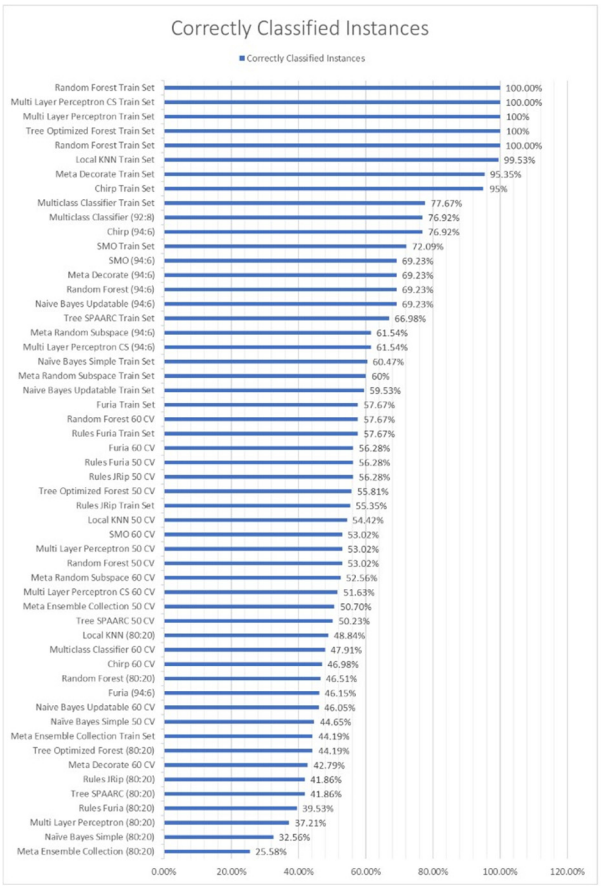
 doi:10.20944/preprints202402.0167.v1

Table 10. Summary of machine learning results.

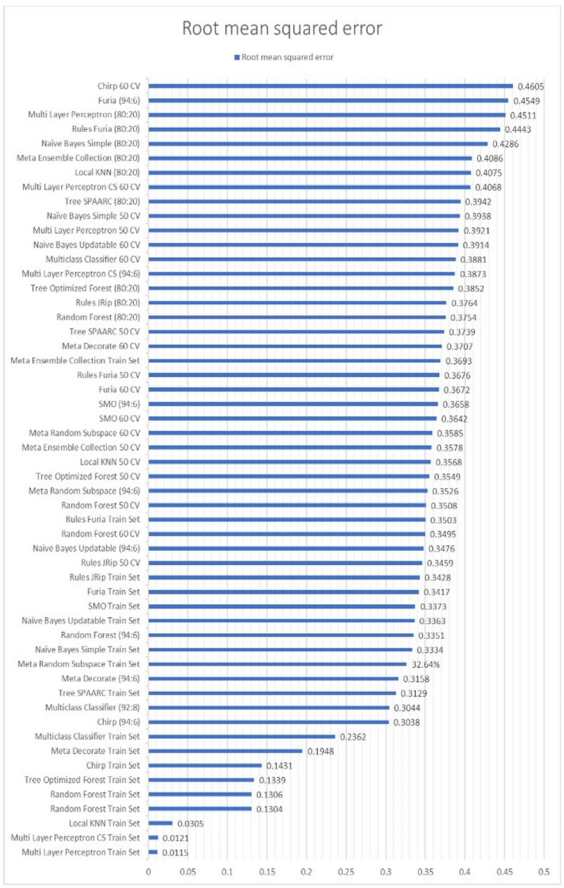
ML Method	Tree SPAARC			Random Forest			Tree Optimized Forest		
Test Types	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**
CCI	41.86%	50.23%	66.98%	46.51%	53.02%	100.00%	44.19%	55.81%	100%
ICI	58.14%	49.77%	33.02%	53.49%	46.98%	0.00%	55.81%	44.19%	0%
RMSE	0.3942	0.3739	0.3129	0.3754	0.3508	0.1304	0.3852	0.3549	0.1339
Total Instances	43	215	215	43	215	215	43	215	215
Precision	N/A	0.451	N/A	0.561	0.503	1	0.617	0.557	1
Recall	0.49	0.502	0.67	0.465	0.53	1	0.442	0.558	1
ML Method	**Local KNN**			**Multi Layer Perceptron**			**Naïve Bayes Simple**		
Test Types	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**
CCI	48.84%	54.42%	99.53%	37.21%	53.02%	100%	32.56%	44.65%	60.47%
ICI	51.16%	45.58%	0.47%	62.79%	46.98%	0%	67.44%	55.35%	39.53%
RMSE	0.4075	0.3568	0.0305	0.4511	0.3921	0.0115	0.4286	0.3938	0.3334
Total Instances	43	215	215	43	215	215	43	215	215
Precision	N/A	0.525	0.995	0.419	0.519	1	0.356	0.446	0.637
Recall	0.488	0.544	0.995	0.372	0.53	1	0.326	0.447	0.605
ML Method	**Meta Ensemble Collection**			**Rules JRip**			**Rules Furia**		
Test Types	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**	**80:20**	**50 CV**	**Train Set**
CCI	25.58%	50.70%	44.19%	41.86%	56.28%	55.35%	39.53%	56.28%	57.67%
ICI	74.42%	49.30%	55.81%	58.14%	43.72%	44.65%	60.47%	43.72%	42.33%
RMSE	0.4086	0.3578	0.3693	0.3764	0.3459	0.3428	0.4443	0.3676	0.3503
Total Instances	43	215	215	43	215	215	43	215	215
Precision	N/A	0.475	N/A	N/A	0.657	N/A	N/A	0.644	N/A
Recall	0.256	0.507	0.442	0.419	0.563	0.553	0.395	0.563	0.577
ML Method	**Naive Bayes Updatable**			**Multi Layer Perceptron CS**			**Meta Random Subspace**		
Test Types	**94:6**	**60 CV**	**Train Set**	**94:6**	**60 CV**	**Train Set**	**94:6**	**60 CV**	**Train Set**
CCI	69.23%	46.05%	59.53%	61.54%	51.63%	100.00%	61.54%	52.56%	60%
ICI	30.77%	53.95%	40.47%	38.46%	48.37%	0.00%	38.46%	47.44%	39.53%
RMSE	0.3476	0.3914	0.3363	0.3873	0.4068	0.0121	0.3526	0.3585	32.64%
Total Instances	13	215	215	13	215	215	13	215	215
Precision	N/A	0.493	0.635	N/A	0.503	1	N/A	0.612	N/A
Recall	0.692	0.46	0.595	0.615	0.516	1	0.615	0.526	0.605
ML Method	**Random Forest**			**Chirp**			**Multiclass Classifier**		
Test Types	**94:6**	**60 CV**	**Train Set**	**94:6**	**60 CV**	**Train Set**	**92:8**	**60 CV**	**Train Set**
CCI	69.23%	57.67%	100.00%	76.92%	46.98%	95%	76.92%	47.91%	77.67%
ICI	30.77%	42.33%	0.00%	23.08%	53.02%	5%	23.08%	52.09%	22.33%
RMSE	0.3351	0.3495	0.1306	0.3038	0.4605	0.1431	0.3044	0.3881	0.2362
Total Instances	13	215	215	13	215	215	13	215	215
Precision	**N/A**	**0.568**	**1**	**N/A**	**0.449**	**0.95**	**0.791**	**0.502**	**0.778**
Recall	**0.962**	**0.577**	**1**	**0.769**	**0.47**	**0.949**	**0.769**	**0.479**	**0.777**
ML Method	**Meta Decorate**			**SMO**			**Furia**		
Test Types	**94:6**	**60 CV**	**Train Set**	**94:6**	**60 CV**	**Train Set**	**94:6**	**60 CV**	**Train Set**
CCI	69.23%	42.79%	95.35%	69.23%	53.02%	72.09%	46.15%	56.28%	57.67%
ICI	30.77%	57.21%	4.65%	30.77%	46.98%	27.91%	53.85%	43.72%	42.33%
RMSE	0.3158	0.3707	0.1948	0.3658	0.3642	0.3373	0.4549	0.3672	0.3417
Total Instances	13	215	215	13	215	215	13	215	215
Precision	N/A	0.435	0.955	N/A	0.519	0.75	N/A	N/A	N/A
Recall	0.692	0.428	0.953	0.692	0.53	0.721	0.462	0.563	0.577

[1] CCI stands for Correctly Classified Instances. [2] ICI stands for Incorrectly Classified Instances.

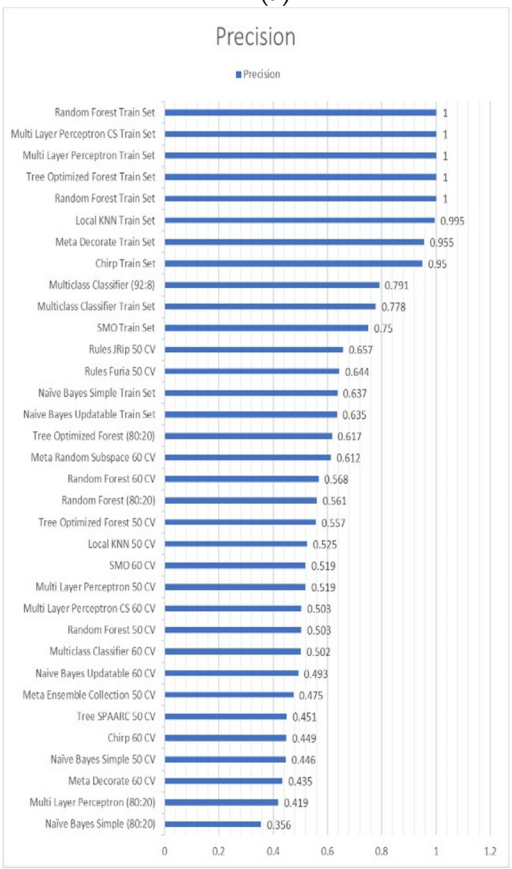
 doi:10.20944/preprints202402.0167.v1



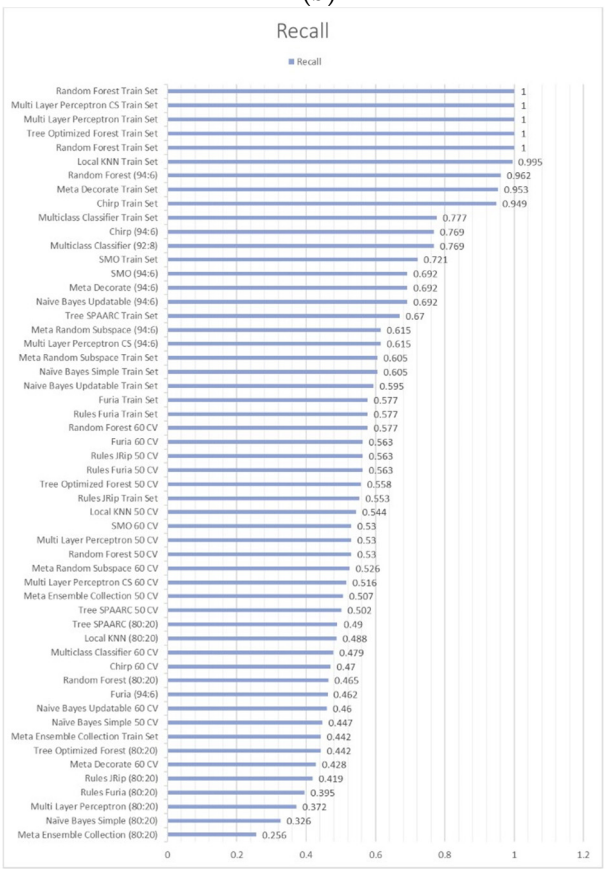
(**a**)



(**b**)



(**c**)



(**d**)

 doi:10.20944/preprints202402.0167.v1

Figure 7. Accuracy evaluation measurement for 15 selected attributes using 15 state-of -the-art machine learning models. The measurement we use includes: (**a**) Correctly classified instances; (**b**) RMSE; (**c**) Precision; (**d**) Recall.

For an overview, we explore the JRip ML model in detail in this section. According to William et al. [31], JRip is an improvement to IREP in terms of a novel heuristic for selecting when to stop adding rules to a rule set and a post pass that "optimizes" a rule set in an attempt to more closely approaches conventional (i.e., non-incremental) reduced error pruning. The measure used to guide pruning is clearly responsible for the occasional failure of IREP to converge with an increasing number of samples; hence, they improved the IREP metric with the formula, yielding better intuition and befitting behavior.

Figure 7 shows that the best-performing ML model is the Multiclass Classifier, with 76.92% correctly classified instances with 92% training data and 8% testing data, and the least performing ML model is the meta-ensemble collection, with 25.58% correctly classified instances with 80% training data and 20% testing data.

The model with the least error is the best-performing model; similarly, the better model has a lower error or RMSE value. The best RMSE from our experimental results with 18 ML models and 15 and 16 attributes was obtained by Chirp (0.3038) using 94% training data split and 6% testing data, as shown in Figure 7b. The worst RMSE with the largest error is Chirp with 60-fold cross validation.

Precision is the ratio of true positives to all the positives. In our problem statement, it would be the proportion of star ratings accurately identified as subjective star rating MOS out of all participants who really had it. The higher the precision value, the better the accuracy of the ML model. As shown in Figure 7c, the best precision value was obtained by the Multiclass Classifier (0.791), and the worst precision value was obtained by Naïve Bayes Simple (0.356).

Recall was the next measurement metric used, as shown in Figure 7d. This is a measure of how well our model identifies the true positives. Thus, recall signifies the number of star ratings that are accurately recognized as real star ratings. In our study, the best recall was obtained by random forest (0.962) with 94% training data and 6% testing data, and the worst recall was obtained by meta-ensemble classification (0.256) with 80% training data and 20% testing data.

The information obtained from the questionnaires is presented in Tables 5 and 10. An analysis of the questionnaires and partial experimental results showed that:

- Massive intense ads may impact QoE and increase ITU results (i.e., higher bitrate, frame rate, and resolution); this does not signify that star review from participants will be high.
- Possible QoE IFs: For acceptable general video content, the advertisement factors, ad length, number of ads, ad location, related ad with content, repeated ad, and maximum ad acceptance are the QoE IFs.
- The most impaired QoE was in mid-roll and unskippable ads (approximately 36.2% and 31.9%, respectively). The QoE that the user can accept to watch an ad of less than 10 seconds is approximately 41.67.

To obtain better MOS recommendations to extend the ITU P.1203 standard, we extracted features to be fed into several ML models. The ML that we employed to predict MOS by considering features such as FER, ITU P.1203, and advertisement results are tree SPAARC, Random Forest, Tree Optimized Forest, Local KNN, Tree Optimized Forest, Multilayer Perceptron, Naive Bayes Simple, Meta Ensemble Collection, Rules JRip and Rules Furia, Naive Bayes Updatable, Multi-Layer Perceptron, meta-decorate, and SMO. We tested all ML models using 80% training data and 20% testing data, 92% training data and 8% testing data, 94% training data and 6% testing data, 100% training data, and 50- and 60-fold cross validations to determine the accuracy of each ML. The experimental results show that class overfitting bias occurs when using 100% training data (Table 10); we excluded the use of 100% training data. Rules JRip and Furia afforded the highest values for correctly classified instances. Moreover, Rules JRip afforded the highest precision recall.

The ITU P.1203 results show that it can correctly predict only 86 of 216 instances or 39.8% of correctly classified instances. The classification accuracy can be improved by considering the FER,

 doi:10.20944/preprints202402.0167.v1

video metadata, and advertisement data, as presented in Table 9. Only the best measurement values were obtained from all testing types. All ML values were obtained through trial and error by sequentially tuning each model parameter and hyperparameter. Hence, we performed hundreds of trainings, testing, and validation steps. All experiments related to ML were performed using Weka version 3.8.6 on Windows 11 Pro, 21H2, Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz.

From the questionnaire, we can obtain several pieces of information that can be seen in Tables 3 and 5. Analysis from questionnaire and partial experimental results that can be drawn:

- Massive intense ads may impact QoE Even ITU results is high (bitrate, framerate, and resolution are high), it does not mean that star review from participants will also be high.
- Possible QoE influence factors: Advertisement After the acceptable general video content, it can be advertisement factors: ad length, number of ads, ad location, related ad with content, repeated ad, and maximum ad acceptance.
- The most impaired QoE was in the mid-long ad and unskippable ads at approximately 36.2% and 31.9%, respectively. The limitation that a user can accept to watch an ad is less than 10 s (approximately 41.67%).

To answer the next research questions on how to provide better recommendation on MOS to extend the ITU P.1203 standard, we extracted some features to be fed into several ML models.

The ITU P.1203 result can only predict 86 instances correctly from 216 instances overall or 39.8% correctly classified instances. From this experiment, including 43 participants, 216 instances, and 16 attributes, we can improve 37.12% of the correctly classified class from ITU-T P.1203 standard results by considering FER, video metadata, and advertisement data, as shown in Table 10. From all testing types, we take only the best value for all measurements. All ML values were obtained by trial and error using a one-by-one ML model and by tuning each of the model parameters and hyperparameters. Hence, we have attempted hundreds of trainings, testing, changing types of testing, and validation steps.

The comparison between ITU-T P.1203 results, face emotion recognition results, and our proposed method results can be seen in Figure 8 and Table C1 in Appendix C.

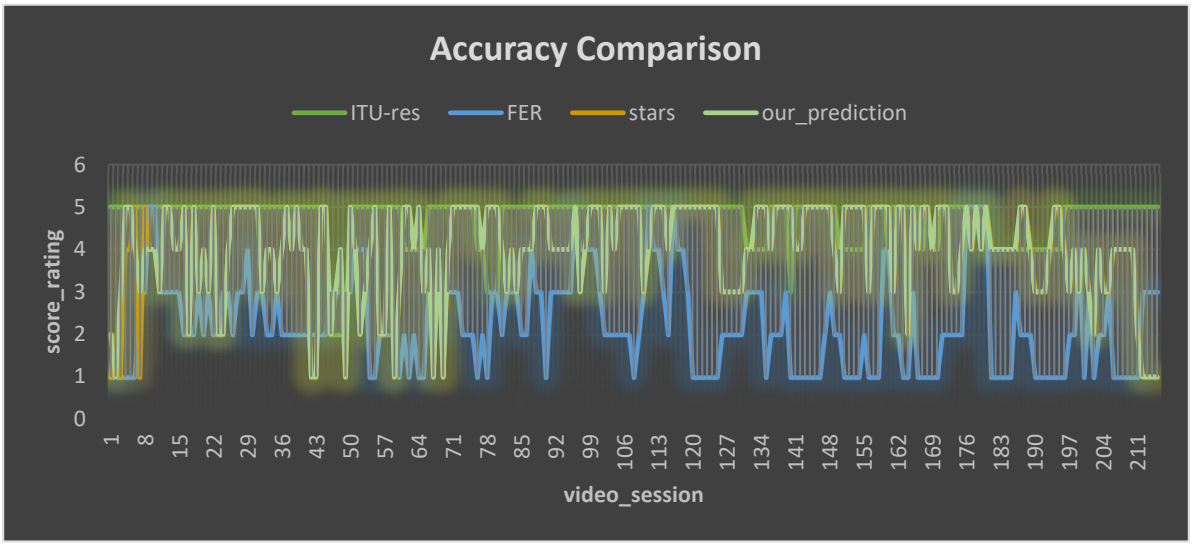


Figure 8. Comparison graph between star reviews, ITU-T P. 1203 results, and FER results.

5. Discussions and Future Directions

We discussed more related matters and some important details that we believe they are important in years to come in this section.

 doi:10.20944/preprints202402.0167.v1

1. Real live FER system: We have tried the real time live system face emotion recognition in the wild and it works effectively although the emotion results do not drive the shape of traffic yet. Our proposed method uses emotion-aware advertisement insertion. Shaping the traffic based on emotion to improve QoE will be our future work.
2. Computational complexity of the system: the computational complexity of the proposed method is dominated by the process of facial emotion recognition (FER). The FER process involves several steps, including: first, face detection: This step involves identifying the location of the face in the video frame. The computational complexity of face detection depends on the DeepFace algorithm. The complexity is $O(n)$, where n is the number of pixels in the video frame. Second, feature extraction: This step involves extracting features from the face. The computational complexity of feature extraction depends on the specific feature used. Therefore, the complexity of the compound becomes O(f), where f is the number of extracted features. Third, motion classification: This step involves classifying the extracted features into one of seven basic emotions (happiness, sadness, anger, fear, surprise, disgust, and neutrality). The computational complexity of emotion classification depends on the specific classifier used. However, in general, it can be thought of as $O(c)$, where c is the sum of emotion classes.
3. Therefore, the overall computational complexity of the FER process can be thought of as $O(n \cdot f \cdot c)$. In the context of the proposed method, the FER process is carried out on each frame of the video. Therefore, the overall computational complexity of the proposed method is $O(T \cdot n \cdot f \cdot c)$, where T is the total number of video frames. For a video that is 30 seconds long and has a frame rate of 30 frames per second, then $T = 30 \times 30 = 900$. If the video frame is 640 × 480 pixels, then $n = 640 \times 480 = 307200$. For $f = 100$ features are extracted from each face, and c = 7 emotions are classified, then the overall computational complexity of the proposed method is $O(900 \cdot 307200 \cdot 100 \cdot 7) = 1.6 \times 1012$. This is a relatively high computational complexity. However, it is important to note that the FER process can be parallelized with the GPU to significantly reduce the computational costs of the proposed method. In addition, it is important to note that the proposed method is used only to select the most relevant ads for a particular user. Once the most relevant ad is selected, including the place, type, time of the ad. Then these types of ads can be provided to users without the need for further facial emotion recognition. Therefore, the overall computational impact of the proposed method is relatively small.
4. Theoretical analysis of our proposed model: for the theoretical analysis of our proposed model, we can see the explanation as follows. The proposed Machine Learning (ML) model for Video Quality of Experience (QoE) inference, incorporating face emotion recognition, user feedback on ad insertion, and network conditions, was evaluated on a dataset of 50 recorded video streaming sessions. This dataset included face videos of viewers, network traffic logs, user feedback on ad insertion, and subjective QoE scores. The model's accuracy was compared against two baseline models: one utilizing only network conditions for QoE inference, ITU-T P.1203 and another employing solely user feedback on ad insertion. The proposed model consistently achieved lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) compared to the baseline models, indicating superior accuracy in inferring QoE. It can be seen in Figure. 8 and Table 9.
5. Qualitative analysis revealed the model's sensitivity to facial expressions of viewers, particularly joy, surprise, and frustration, known indicators of positive and negative QoE, respectively. It also learned to identify ad placements perceived as disruptive by users, adjusting its QoE predictions accordingly. Moreover, the model effectively utilized network bandwidth as a critical indicator of potential rebuffering and stalling, which negatively impacts QoE. These experimental results convincingly demonstrate the effectiveness of the proposed ML model in accurately inferring video QoE. Its ability to integrate face emotion recognition, user feedback on ad insertion, and network conditions provides a comprehensive understanding of QoE, holding significant promise for improving user satisfaction and network performance in video streaming systems.
6. Our hypothesis by mapping the extracted emotion to ACR and MOS are based on two studies by Porcu et al. [40] and Martinez-Caro & Cano [52]. Porcu et al. [40] analyzing the facial expression and gaze direction achieved 93.9% accuracy leveraging k-NN classifier which investigates the possibility to estimate perceived QoE by using face emotion and gaze

 doi:10.20944/preprints202402.0167.v1

movement. Moreover, the work by Martinez-Caro & Cano [52] utilizing the ITU-T P.1203 model to estimate MOS value and uses variational algorithms to predict QoE to give insight on emotional impact of video quality.

6. Conclusions

Based on the experimental results and all procedures performed, we can conclude that QoE is influenced by the existence of advertisements. Advertisement length, frequency, number of repetitions, relation to content, and location impact QoE. The ITU and FER results alone cannot guarantee accurate QoE estimation. The relation between ITU results, FER results, questionnaire, and star review could provide guidance for advertisement placement in the future.

- Users get more annoyed when an advertisement is placed in the middle of the content.
- The maximum tolerable advertisement length is 10 s.
- The most annoying advertisement is an unskippable one.
- The most degraded QoE was obtained for a mid-roll advertisement.

Furthermore, according to the ITU P.1203 results, only 86 of the 216 occurrences were properly forecasted, and 39.8% of the instances were correctly categorized. Rule Chirp, on the other hand, accurately predicted 121 of 216 cases (76.92%). In this experiment with 43 participants, 216 occurrences s, and 15 characteristics, the accuracy of the ITU-T P.1203 standard was improved by considering the FER, ITU-T P.1203 result, and advertisement data. We propose an answer to the research question of how to provide network administrators with insight or prediction of QoE in the end-to-end encrypted content in their network with an accuracy of 76.92% and recall of 0.962 by considering face emotion, advertisement data, and video statistics. A network administrator may react to any impairment discovered using this inference, and accordingly QoE may be immediately improved.

Author Contributions: All the authors contributed equally to this work.

Funding: This research work is supported by UAEU Grant: 31T102-UPAR-1-2017.

Informed Consent Statement: Written informed consent has been obtained from the participants to publish this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1 lists all the abbreviations used in this article.

Table A1. Summary of machine learning results.

Abbreviation	Stands For
CCI	Correctly Classified Instances
CNN	Convolutional neural network
CV	Cross Validation
DL	Deep learning
FER	Face Emotion Recognition
HAS	HTTP Adaptive Streaming
HTTP	Hypertext Transfer Protocol
ICI	Incorrectly Classified Instances
ITU-T	International Telecommunication Union - Telecommunication
Mid-roll	Video advertisement at the middle of content playback
MSE	Mean squared error
Post-roll	Video advertisement at the end of content playback
Pre-roll	Video advertisement before content playback started
QoE	Quality of Experience

 doi:10.20944/preprints202402.0167.v1

QoS	Quality of Service
UE	User Experience
Weka	Waikato Environment for Knowledge Analysis
MSE	Mean squared error

Appendix B

In this section we list some questions asked in the questionnaire. The questionnaire results in detail we planned that it will be elaborated in another journal.

1. What do you think about the frequency of advertisement in pre-roll
2. What do you think about the frequency of advertisement in mid-roll
3. What do you think about the frequency of advertisement in post-roll
4. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Relaxed - (B) Stimulated]
5. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Calm - (B) Excited]
6. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Sluggish - (B) Frenzied]
7. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Dull - (B) Jittery]
8. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Sleepy - (B) Wide Awake]
9. How bodily relaxed / aroused are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Unaroused - (B) Aroused]
10. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Relaxed - (B) Stimulated]
11. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Calm - (B) Excited]
12. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Sluggish - (B) Frenzied]
13. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Dull - (B) Jittery]
14. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Sleepy - (B) Wide Awake]
15. How bodily relaxed / aroused are you after watching second video (Title: First Underwater Farm)? [(A) Unaroused - (B) Aroused]
16. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Relaxed - (B) Stimulated]
17. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Calm - (B) Excited]
18. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Sluggish - (B) Frenzied]
19. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Dull - (B) Jittery]
20. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Sleepy - (B) Wide Awake]
21. How bodily relaxed / aroused are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Unaroused - (B) Aroused]
22. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Relaxed - (B) Stimulated]
23. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Calm - (B) Excited]
24. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Sluggish - (B) Frenzied]

 doi:10.20944/preprints202402.0167.v1

25. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Dull - (B) Jittery]
26. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Sleepy - (B) Wide Awake]
27. How bodily relaxed / aroused are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Unaroused - (B) Aroused]
28. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Relaxed - (B) Stimulated]
29. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Calm - (B) Excited]
30. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Sluggish - (B) Frenzied]
31. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Dull - (B) Jittery]
32. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Sleepy - (B) Wide Awake]
33. How bodily relaxed / aroused are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Unaroused - (B) Aroused]
34. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Controlled - (B) Controlling]
35. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Influenced - (B) Influential]
36. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Cared for - (B) In control]
37. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Awed - (B) Important]
38. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Submissive - (B) Dominant]
39. How emotionally controlled / uncontrolled are you after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Guided - (B) Autonomous]
40. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Controlled - (B) Controlling]
41. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Influenced - (B) Influential]
42. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Cared for - (B) In control]
43. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Awed - (B) Important]
44. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Submissive - (B) Dominant]
45. How emotionally controlled / uncontrolled are you after watching second video (Title: First Underwater Farm)? [(A) Guided - (B) Autonomous]
46. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Controlled - (B) Controlling]
47. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Influenced - (B) Influential]
48. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Cared for - (B) In control]
49. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Awed - (B) Important]
50. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Submissive - (B) Dominant]
51. How emotionally controlled / uncontrolled are you after watching third video (Title: Most Beautiful Building In The World)? [(A) Guided - (B) Autonomous]

 doi:10.20944/preprints202402.0167.v1

52. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Controlled - (B) Controlling]
53. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Influenced - (B) Influential]
54. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Cared for - (B) In control]
55. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Awed - (B) Important]
56. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Submissive - (B) Dominant]
57. How emotionally controlled / uncontrolled are you after watching fourth video (Title: This is Made of...PEE?!)? [(A) Guided - (B) Autonomous]
58. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Controlled - (B) Controlling]
59. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Influenced - (B) Influential]
60. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Cared for - (B) In control]
61. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Awed - (B) Important]
62. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Submissive - (B) Dominant]
63. How emotionally controlled / uncontrolled are you after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Guided - (B) Autonomous]
64. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Unhappy - (B) Happy]
65. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Annoyed - (B) Pleased]
66. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Unsatisfied - (B) Satisfied]
67. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Melancholic - (B) Contented]
68. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Despairing - (B) Hopeful]
69. How pleasant / unpleasant do you feel after watching first video (Title: How This Guy Found a Stolen Car!)? [(A) Bored - (B) Relaxed]
70. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Unhappy - (B) Happy]
71. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Annoyed - (B) Pleased]
72. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Unsatisfied - (B) Satisfied]
73. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Melancholic - (B) Contented]
74. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Despairing - (B) Hopeful]
75. How pleasant / unpleasant do you feel after watching second video (Title: First Underwater Farm)? [(A) Bored - (B) Relaxed]
76. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Unhappy - (B) Happy]
77. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Annoyed - (B) Pleased]
78. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Unsatisfied - (B) Satisfied]

 doi:10.20944/preprints202402.0167.v1

79. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Melancholic - (B) Contented]
80. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Despairing - (B) Hopeful]
81. How pleasant / unpleasant do you feel after watching third video (Title: Most Beautiful Building In The World)? [(A) Bored - (B) Relaxed]
82. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Unhappy - (B) Happy]
83. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Annoyed - (B) Pleased]
84. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Unsatisfied - (B) Satisfied]
85. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Melancholic - (B) Contented]
86. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Despairing - (B) Hopeful]
87. How pleasant / unpleasant do you feel after watching fourth video (Title: This is Made of...PEE?!)? [(A) Bored - (B) Relaxed]
88. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Unhappy - (B) Happy]
89. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Annoyed - (B) Pleased]
90. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Unsatisfied - (B) Satisfied]
91. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Melancholic - (B) Contented]
92. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Despairing - (B) Hopeful]
93. How pleasant / unpleasant do you feel after watching fifth video (Title: The Most Unexplored Place In The World)? [(A) Bored - (B) Relaxed]
94. Reduction of service experience due to pre-roll advertisement
95. What do you think about the reduction of service experience due to mid-roll advertisement
96. What do you think about the reduction of service experience due to post-roll advertisement
97. How do you feel about the annoyance due to pre-roll advertisement
98. How do you feel about the annoyance due to mid-roll advertisement
99. How do you feel about the annoyance due to post-roll advertisement
100. What is your opinion about maximum acceptable advertisement length period
101. Please reorder from the most annoying to the least:.
    1. Many repeated ads in 1 point of time in mid-roll.
    2. Single 5 minutes ads long in mid-roll.
    3. In 5 minutes, video content, every 1 minute there is one repeated ad.
    4. Same ads repeatedly in the pre-roll, mid-roll, post-roll.
    5. There are no skippable ads.

Appendix C

In this section, we list the table of results that built Figure 8. This table compare ITU-T P.1203 results, Face emotion recognition results, our proposed method results and as a ground truth the star rating given by participants.

Table C1. Comparison results of ITU-T P.1203 result, FER, star and prediction.

ITU-res	FER	Star (ground truth)	Our Prediction
5	**1**	**1**	**2**
5	**1**	**2**	**1**
5	**1**	**1**	**3**

 doi:10.20944/preprints202402.0167.v1

5	1	4	5
5	1	4	5
5	1	5	4
5	3	1	3
5	3	5	4
5	5	4	4
5	5	4	4
5	3	3	3
5	3	5	5
5	3	5	5
5	3	4	4
5	3	4	4
5	2	5	5
5	2	2	2
5	2	5	5
5	3	3	3
5	2	4	4
5	3	3	3
5	2	5	5
5	2	2	2
5	3	2	2
5	3	4	4
5	2	5	5
5	3	5	5
5	3	5	5
5	4	5	5
5	2	5	5
5	3	5	5
5	3	3	3
5	2	4	4
5	2	4	4
5	3	3	3
5	2	4	4
5	2	5	5
5	2	4	4
5	2	5	5
5	2	4	4
5	2	4	4
5	2	1	1
5	2	1	1
5	2	5	5
5	2	5	5
2	3	2	2
2	3	3	3
2	3	4	4
2	3	1	1
5	3	5	5
5	4	4	4
5	4	2	2

 doi:10.20944/preprints202402.0167.v1

5	4	3	3
5	1	4	4
5	1	5	5
5	2	2	2
5	2	2	2
5	2	5	5
5	2	1	1
5	1	3	3
4	2	5	5
4	1	5	5
4	2	4	4
4	1	5	5
4	1	3	3
5	3	4	4
5	3	1	1
5	3	4	4
5	3	1	1
5	3	4	4
5	3	5	5
5	3	5	5
5	2	5	5
5	2	5	5
5	2	5	5
4	1	5	5
4	2	4	4
3	1	5	5
3	3	5	5
3	3	5	5
5	2	4	4
5	3	3	3
5	3	4	4
5	3	3	3
5	2	4	4
5	2	4	4
5	4	4	4
5	3	5	5
5	3	5	5
5	1	5	5
5	3	4	4
5	3	4	4
5	3	4	4
5	3	4	4
5	3	4	4
5	5	5	5
5	4	3	3
5	4	4	4
5	4	5	5
5	4	5	5
5	3	5	5

 doi:10.20944/preprints202402.0167.v1

5	2	3	3
5	2	5	5
5	2	4	4
5	2	5	5
5	2	5	5
5	2	5	5
5	1	5	5
5	2	5	5
5	3	3	3
5	5	4	4
5	4	5	5
5	4	5	5
5	3	5	5
5	2	4	4
5	5	5	5
5	4	5	5
5	4	5	5
5	3	5	5
5	1	5	5
5	1	5	5
5	1	5	5
5	1	5	5
5	1	5	5
5	1	5	5
5	2	3	3
5	2	3	3
5	2	3	3
5	2	3	3
5	2	3	3
4	3	4	4
4	3	4	4
4	3	5	5
4	3	5	5
4	1	5	5
5	2	4	4
5	2	5	5
5	2	5	5
5	3	5	5
3	1	5	5
5	1	4	4
5	1	4	4
5	1	5	5
5	1	5	5
5	1	5	5
5	1	5	5
5	2	5	5
5	3	5	5
5	2	4	4
4	2	3	3

 doi:10.20944/preprints202402.0167.v1

4	1	5	5
4	1	5	5
4	1	5	5
4	1	5	5
4	2	5	5
5	1	3	3
5	1	5	5
5	1	5	5
5	4	5	5
4	4	5	5
5	2	3	3
5	2	5	5
5	1	5	5
5	1	2	2
5	3	5	5
5	1	4	4
4	1	5	5
5	1	5	5
4	1	4	4
4	1	5	5
5	2	5	5
5	2	5	5
5	2	3	3
5	2	3	3
5	2	4	4
5	5	5	5
5	5	4	4
5	5	5	5
5	5	4	4
5	5	5	5
4	1	4	4
4	1	4	4
4	1	4	4
4	1	4	4
4	1	4	4
4	3	4	4
4	2	5	5
4	2	5	5
4	2	4	4
4	1	3	3
4	1	3	3
4	1	3	3
4	1	4	4
4	1	5	5
4	1	5	5
5	1	4	4
5	2	3	3
5	2	4	4
5	4	3	3

 doi:10.20944/preprints202402.0167.v1

5	**1**	**4**	**4**
5	**3**	**3**	**3**
5	**1**	**2**	**2**
5	**2**	**4**	**4**
5	**2**	**3**	**3**
5	**3**	**3**	**3**
5	**1**	**4**	**4**
5	**1**	**4**	**4**
5	**1**	**4**	**4**
5	**1**	**4**	**4**
5	**1**	**4**	**4**
5	**1**	**2**	**2**
5	**3**	**1**	**1**
5	**3**	**1**	**1**
5	**3**	**1**	**1**
5	**3**	**1**	**1**

References

1. Gutterman, C.; et al. Requet: Real-Time Quantitative Detection for Encrypted YouTube Traffic. In Proceedings of the 10th ACM Multimedia System Conference, **2019**. [CrossRef]
2. Izima, O.; de Fréin, R.; Malik, A. A survey of machine learning techniques for video quality prediction from quality of delivery metrics. Electronics 2021, 10 (22), 2851. [CrossRef]
3. Agboma, F.; Liotta, A. "QoE-Aware QoS Management." Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia **2008**. [CrossRef]
4. Streijl, R. C.; Winkler, S.; Hands, D. S. "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives." Multimedia Systems **2016**, 22 (2), 213-227. [CrossRef]
5. Engelke, U.; Darcy, D. P.; Mulliken, G. H.; Bosse, S.; Martini, M. G.; Arndt, S.; Antons, J. N.; Chan, K. Y.; Ramzan, N.; Brunnström, K. "Psychophysiology-Based QoE Assessment: A Survey." IEEE Journal of Selected Topics in Signal Processing **2016**, 11 (1), 6-21. [CrossRef]
6. Raake, A.; et al. "A Bitstream-Based, Scalable Video-Quality Model for HTTP Adaptive Streaming: ITU-T P. 1203.1." **2017** Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017. [CrossRef]
7. Garcia, M.-N.; Dytko, D.; Raake, A. Quality Impact Due to Initial Loading, Stalling, and Video Bitrate in Progressive Download Video Services. 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX) **2014**, pp. 129–134. [CrossRef]
8. Garcia, M. N.; Dytko, D.; Raake, A. "Quality Impact Due to Initial Loading, Stalling, and Video Bitrate in Progressive Download Video Services." 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX). IEEE, **2014**, 129-134. [CrossRef]
9. Pereira, R.; Pereira, E. G. "Dynamic Adaptive Streaming over HTTP and Progressive Download: Comparative Considerations." 2014 28th International Conference on Advanced Information Networking and Applications Workshops. IEEE, **2014**, 905-909. [CrossRef]
10. Sackl, A.; Zwickl, P.; Reichl, P. The trouble with choice: An empirical study to investigate the influence of charging strategies and content selection on QoE. Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013) **2013**, 298–303. [CrossRef]
11. Hoßfeld, T.; Seufert, M.; Hirth, M.; Zinner, T.; Tran-Gia, P.; Schatz, R. Quantification of YouTube QoE via crowdsourcing. **2011** IEEE International Symposium on Multimedia 2011, 494–499. [CrossRef]
12. Oyman, O.; Singh, S. Quality of experience for HTTP adaptive streaming services. IEEE Communications Magazine **2012**, 50 (4), 20–27. [CrossRef]
13. Yao, J.; Kanhere, S. S.; Hossain, I.; Hassan, M. Empirical evaluation of HTTP adaptive streaming under vehicular mobility. International Conference on Research in Networking **2011**, 92–105. [CrossRef]
14. Ghani, R. F.; Ajrash, A. S. Quality of Experience Metric of Streaming Video: A Survey. Iraqi Journal of Science **2018**, 59 (3B), 1531–1537. [CrossRef]

15. Porcu, S.; Floris, A.; Atzori, L. Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction. 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN). IEEE **2019**, pp. 82–87. [CrossRef]
16. Akhshabi, S.; Anantakrishnan, L.; Begen, A. C.; Dovrolis, C. What happens when adaptive streaming players compete for bandwidth? Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video **2012**, 9–14. [CrossRef]
17. Zinner, T.; Hossfeld, T.; Minhas, T. N.; Fiedler, M. Controlled vs. uncontrolled degradations of QoE: The provisioning-delivery hysteresis in case of video. EuroITV 2010 Workshop: Quality of Experience for Multimedia Content Sharing **2010**. [CrossRef]
18. Cohen, W. W. Fast Effective Rule Induction. In Machine Learning Proceedings 1995. Elsevier **1995**, pp. 115–123. [CrossRef]
19. Landis, J. R.; Koch, G. G. An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics **1977**, pp. 363–374. [CrossRef]
20. Bermudez, H.-F.; et al. Live Video-Streaming Evaluation Using the ITU-T P. 1203 QoE Model in LTE Networks. Computer Network **2019**, 165, 106967. [CrossRef]
21. Wang, W.; Lu, Y. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. IOP Conference Series: Materials Science and Engineering **2018**, 324 (1), 012049. [CrossRef]
22. Seshadrinathan, K.; Soundararajan, R.; Bovik, A. C. Study of Subjective and Objective Quality Assessment of Video. IEEE Trans. Image Process. **2010**, 19 (6), 1427–1441. [CrossRef]
23. Amour, L.; Lamine, et al. "An Improved QoE Estimation Method Based on QoS and Affective Computing." 2018 Int. Symp. Program. Syst. **2018**. [CrossRef]
24. Bhattacharya, A.; Wu, W.; Yang, Z. "Quality of Experience Evaluation of Voice Communication: An Affect-Based Approach." Hum.-Centric Comput. Inf. Sci. **2012**, 2 (1), 1-18. [CrossRef]
25. Callet, P.; Möller, S.; Perkis, A. "Qualinet White Paper on Definitions of Quality of Experience (2012)." Eur. Network Qual. Exp. Multimedia Syst. and Serv. **2013**.
26. Porcu, S.; et al. "Emotional Impact of Video Quality: Self-Assessment and Facial Expression Recognition." 2019 11th Int. Conf. Qual. Multimedia Exp. (QoMEX). **2019**. [CrossRef]
27. Antons, J.-N.; et al. "Analyzing Speech Quality Perception Using Electroencephalography." IEEE J. Sel. Top. Signal Process. **2012**, 6 (6), 721-731. [CrossRef]
28. Kroupi, E.; et al. "EEG Correlates During Video Quality Perception." 2014 22nd Eur. Signal Process. Conf. (EUSIPCO). **2014**.
29. Arndt, S.; et al. "Using Electroencephalography to Analyze Sleepiness Due to Low-Quality Audiovisual Stimuli." Signal Process. Image Commun. **2016**, 42, 120-129. [CrossRef]
30. Arndt, S.; et al. "Using Eye-Tracking and Correlates of Brain Activity to Predict Quality Scores." 2014 Sixth Int. Workshop Qual. Multimedia Exp. (QoMEX). **2014**. [CrossRef]
31. Engelke, U.; et al. "Linking Distortion Perception and Visual Saliency in H. 264/AVC Coded Video Containing Packet Loss." Visual Commun. Image Process. 2010, 7744. [CrossRef]
32. Rai, Y.; Le Callet, P. "Do Gaze Disruptions Indicate the Perceived Quality of Nonuniformly Coded Natural Scenes?." Electron. Imaging **2017**, 14, 104-109.
33. Rai, Y.; Barkowsky, M.; Le Callet, P. "Role of Spatio-Temporal Distortions in the Visual Periphery in Disrupting Natural Attention Deployment." Electron. Imaging **2016**, 16, 1-6.
34. Bailenson, J. N.; et al. "Real-time Classification of Evoked Emotions Using Facial Feature Tracking and Physiological Responses." Int. J. Hum.-Comput. Stud. 2008, 66(5), 303-317. [CrossRef]
35. Pereira, R.; Pereira, E. G. "Dynamic Adaptive Streaming Over HTTP and Progressive Download: Comparative Considerations." 28th International Conference on Advanced Information Networking and Applications Workshops **2014**, 905–909. [CrossRef]
36. Robitza, W.; Göring, S.; Raake, A.; Lindegren, D.; Heikkilä, G.; Gustafsson, J.; List, P.; Feiten, B.; Wüstenhagen, U.; Garcia, M. N.; et al. "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P. 1203: Open Databases and Software." Proceedings of the 9th ACM Multimedia Systems Conference 2018, 466-471. [CrossRef]
37. International Telecommunication Union. "Recommendation ITU-T P. 1203.3, Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services over Reliable Transport-Quality Integration Module." **2017**.

38. Raake, A.; Garcia, M. N.; Robitza, W.; List, P.; Göring, S.; Feiten, B. "A Bitstream-Based, Scalable Video-Quality Model for HTTP Adaptive Streaming: ITU-T P. 1203.1." 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, **2017**, 1-6. [CrossRef]
39. Bentaleb, A.; Taani, B.; Begen, A. C.; Timmerer, C.; Zimmermann, R. A survey on bitrate adaptation schemes for streaming media over HTTP. IEEE Communications Surveys Tutorials **2018**, 21 (1), 562–585. [CrossRef]
40. Porcu, S. Estimation of the QoE for Video Streaming Services Based on Facial Expressions and Gaze Direction. **2021**.
41. Roettgers, J. Don't touch that dial: How YouTube is bringing adaptive streaming to mobile, TVs. **2013**.
42. Seufert, M.; Egger, S.; Slanina, M.; Zinner, T.; Hoßfeld, T.; Tran-Gia, P. A survey on quality of experience of HTTP adaptive streaming. IEEE Communications Surveys Tutorials 2014, 17 (1), 469–492. [CrossRef]
43. Barman, N.; Martini, M. G. "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges." IEEE Access **2019**, 7, 30831–30859. [CrossRef]
44. Barakovic´, S.; Skorin-Kapov, L. "Survey of Research on Quality of Experience Modelling for Web Browsing." Quality and User Experience **2017**, 2(1), 1–31. [CrossRef]
45. Cofano, G.; De Cicco, L.; Zinner, T.; Nguyen-Ngoc, A.; Tran-Gia, P.; Mascolo, S. Design and experimental evaluation of network-assisted strategies for HTTP adaptive streaming. Proceedings of the 7th International Conference on Multimedia Systems **2016**, 1–12. [CrossRef]
46. Akhtar, Z.; Falk, T. H. Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey. IEEE Access **2017**, 5, 21 090–21 117. [CrossRef]
47. Zhao, T.; Liu, Q.; Chen, C. W. "QoE in Video Transmission: A User Experience-Driven Strategy." IEEE Communications Surveys Tutorials **2016**, 19 (1), 285-302. [CrossRef]
48. Han, L.; Yu, L. "A Variance Reduction Framework for Stable Feature Selection." Statistical Analysis and Data Mining: The ASA Data Science Journal **2012**, 5(5), 428-445. [CrossRef]
49. Strobl, C.; et al. "Conditional Variable Importance for Random Forests." BMC Bioinformatics **2008**, 9(1).[CrossRef]
50. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. "Permutation Importance: A Corrected Feature Importance Measure." Bioinformatics **2010**, 26(10), 1340-1347. DOI: https://doi.org/10.1093/bioinformatics/btq134 [CrossRef]
51. Menze, B.; et al. "A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data." BMC Bioinformatics **2009**, 10(1). [CrossRef]
52. Martinez-Caro, J.-M.; Cano, M.-D. On the Identification and Prediction of Stalling Events to Improve QoE in Video Streaming. *Electronics* **2021**, *10*, 753. [CrossRef]