



Article

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

Tisa Selma <sup>1,\*</sup> , Mohammad Mehedy Masud <sup>1,\*</sup>, Abdelhak Bentaleb <sup>2</sup> and Saad Harous <sup>3</sup>

<sup>1</sup> College of Information Technology, United Arab Emirates University, Al Ain P.O. Box 15551, United Arab Emirates

<sup>2</sup> Computer Science and Software Engineering, Concordia University, Montreal, QC H3G 1M8, Canada; abdelhak.bentaleb@concordia.ca

<sup>3</sup> Department of Computer Science, University of Sharjah, Sharjah P.O. Box 26666, United Arab Emirates; harous@sharjah.ac.ae

\* Correspondence: 201990100@uaeu.ac.ae (T.S.); m.masud@uaeu.ac.ae (M.M.M.)

**Abstract:** This study introduces an FER-based machine learning framework for real-time QoE assessment in video streaming. This study's aim is to address the challenges posed by end-to-end encryption and video advertisement while enhancing user QoE. Our proposed framework significantly outperforms the base reference, ITU-T P.1203, by up to 37.1% in terms of accuracy and 21.74% after attribute selection. Our study contributes to the field in two ways. First, we offer a promising solution to enhance user satisfaction in video streaming services via real-time user emotion and user feedback integration, providing a more holistic understanding of user experience. Second, high-quality data collection and insights are offered by collecting real data from diverse regions to minimize any potential biases and provide advertisement placement suggestions.

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



**Citation:** Selma, T.; Masud, M.M.; Bentaleb, A.; Harous, S. Inference Analysis of Video Quality of Experience in Relation with Face Emotion, Video Advertisement, and ITU-T P.1203. *Technologies* **2024**, *12*, 62. <https://doi.org/10.3390/technologies12050062>

Academic Editor: Alessandro Tognetti

Received: 28 March 2024

Revised: 20 April 2024

Accepted: 23 April 2024

Published: 3 May 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Video services and related technologies have rapidly advanced in recent decades. Inferring Quality of Experience (QoE) from encrypted data in video streaming apps is a complex task that network service providers must address. All abbreviations are listed in Table A1 and 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 [2,3]. 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 [4]. However, for deteriorated QoS (high jitter, delay, and packet loss), the chances of obtaining adequate QoE are low. In this study, objective QoE refers to QoE that can be calculated without any user feedback, and subjective QoE refers to 1–5 user ratings for certain services or multimedia, also known as the Mean Opinion Score (MOS) [5]. Although psychological behavior, user background, user viewing history, user favorite program, advertisement preferences, and other QoE-influencing factors cannot be controlled, QoS and objective QoE should be prioritized to obtain optimal final QoE. This inference effort attempts to capture what users like by considering advertisement positions, period duration, and recurrence intensity.

The QoS is used to indicate the optimal level of services provided by a multimedia service. Currently, the QoE metric indicates the level of user satisfaction by considering the user's background, psychological–emotional state, and expectancy. The QoS usually depends on several objective metrics, including delay variance, jitter, throughput, and delay.

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 and can be determined based on subjective measurements, usually called the mean opinion score (MOS), in which 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 facial behaviors while watching videos to reduce the cost, time, and effort required to perform an accurate subjective QoE assessment. The major challenge of facial emotion recognition (FER) in video advertising lies in achieving accurate and reliable emotion recognition within a dynamic real-time environment filled with disturbance. Conventional FER systems often struggle with factors like lighting variations, head movements, and occlusions, leading to biased accuracy. Our study effectively addresses this challenge by offering a novel approach that utilizes FER and user feedback alongside advertisement data and ITU-T P.1203 results. This multi-modal approach offers a more comprehensive understanding of user experience, going beyond just facial emotion recognition. While common FER accuracy can range from 30% to 40%, our method achieves a significant improvement with a recall of almost 1. This offers a more sophisticated perception of user emotions and their impacts on QoE within the context of video advertising.

Our study proposes an innovative method utilizing facial emotion recognition to infer QoE, and examines the impact of ads on user experience. We used our own extracted Face Emotion Recognition (FER) datasets, and facial emotion information was extracted from actual observers to train the machine learning models. Our approach aligns with the psychophysiology-based QoE assessment explained by Engelke et al. [6], 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 via factors that may impact QoE, such as advertisements. We chose advertisement as the impairing factor among the other factors because users encounter advertisements in most video streaming sessions. Furthermore, we aim to investigate how and where advertisements can be displayed during video streaming sessions without impairing the end-user QoE. We hypothesized that user QoE would decrease with an increase in the number of advertisements.

We also hypothesized that the better and more stable the video quality and content and the lower the number of advertisements unrelated to content, the better the end-user QoE, and the more satisfied the user will be. The objectives of our study are as follows:

1. To evaluate the effects of video advertisements on QoE—a reliable measurement of QoE is a fundamental step in multimedia communication. Considering all the factors that influence QoE, experiments are important;
2. The experimental results were compared and analyzed based on the ITU-T P.1203 standard and previous studies [7];
3. We propose an accurate machine learning approach that can estimate QoE by considering advertisements and user expressions.

Acceptable end-user QoE is critical because video content and advertisements are widely used. A combination of subjective and objective QoE assessments is required to obtain valid feedback on service performance. The following challenges are encountered 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 for this problem;
2. The ITU-T P.1203 standard QoE measurements did not anticipate significant factors that could lead to inaccurate results. We hope to alleviate this issue using our proposed approach;
3. Excessive video advertisements during a video streaming session may weaken user engagement and negatively affect QoE. We attempted to devise a method in which a viewing session could coexist with several advertisements if a specific threshold was met.

Based on the above research concerns, the following research questions and their corresponding objectives and contributions were obtained.

1. Research Question 1: Leveraging Machine Learning for QoE Inference
  - 1.1. Challenge: Existing methods such as ITU-T P.1203 offer limited accuracy owing to the lack of consideration of user emotions, advertisement data, and real-time video quality conditions.
  - 1.2. Objective: To develop a novel ML approach that integrates ITU-T P.1203 results, facial expressions, and advertisement data to infer video QoE in real time.
  - 1.3. Contribution: The ML approach was proposed and compared with state-of-the-art algorithms, demonstrating its superior accuracy and real-time applicability for video QoE inference.
2. Research Question 2: High-Quality Data Collection and Preparation
  - 2.1. Challenge: Ensure the quality and diversity of the collected data while minimizing potential biases in the participant responses.
  - 2.2. Objective: To collect and prepare a comprehensive and unbiased dataset that is suitable for effective ML model training.
  - 2.3. Contribution:
    - We designed and utilized online platforms to collect data from a large and diverse pool of participants (661 from 114 countries);
    - We developed comprehensive questionnaires (100+ questions) to minimize bias and ensure data accuracy;
    - We implemented data pre-processing techniques (feature extraction and attribute selection) to enhance data quality and model performance;
    - We enriched the model with training data from various sources.
3. Research Question 3: Data Quality Improvement for Machine Learning
  - 3.1. Challenge: Identify the most effective data pre-processing techniques for a specific dataset and model while balancing data quality and quantity, handling outliers, and addressing potential inconsistencies.
  - 3.2. Objective: Enhance data quality to improve ML model performance and generalizability.
  - 3.3. Contribution:
    - Implementing appropriate data pre-processing techniques to address the identified challenges;
    - Utilizing various data sources to enrich the model and enhance its accuracy and generalizability;
    - Iterative ML experiments and adjustments were performed to optimize model training while considering the challenges encountered.
4. Research Question 4: Balancing Advertisement Placement and User Experience
  - 4.1. Challenge: Understanding diverse cultural preferences and user behaviors regarding advertisements across different regions and demographics, while balancing the maximizing of advertisement effectiveness and preserving user engagement and video QoE.
  - 4.2. Objective: Investigate optimal advertisement strategies that balance advertiser budget with user experience.
  - 4.3. Contribution:
    - Conducting surveys to gather diverse cultural and user-centric insights into acceptable advertising practices;
    - Providing recommendations for optimizing advertisement duration and placement strategies to balance budget, user engagement, and video QoE, considering the identified cultural and behavioral factors.

## 2. Background and Related Work

Several OTT services use end-to-end encryption to enhance user privacy and security. However, this encryption utilization may limit the network operator's ability to observe and rectify quality degradation by employing certain functions such as quality-of-service (QoS) provisioning. Currently, several approaches, such as traditional machine-learning (ML)-based and session-based models, are used for video QoE inference in encrypted applications [8]. If 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 inference tasks owing to recently advanced encryption technologies. On-the-fly decryption is unfeasible because of the rapid development of encryption technologies.

In the rapidly evolving landscape of video streaming services, the increasing popularity of end-to-end encryption presents challenges for network administrators striving to uphold the network performance and user experience. The intrusion of video ads into the streaming ecosystem poses the risk of traffic congestion and diminished Quality of Experience (QoE) for users. Traditional algorithms designed to enhance QoE encounter limitations owing to 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. By leveraging open-access Face Emotion Recognition (FER) datasets, facial emotion information was extracted from actual observers to train the machine learning models by leveraging open-access FER datasets. This approach aligns with psychophysiology-based QoE assessment trends, as Engelke et al. (2016) [6] highlighted, acknowledging the importance of understanding the emotional aspects of user experience.

To validate our proposed approach, participants were asked to watch an advertisement and video, provide a rating, and then use the assessment as a basis for comparison, training, testing, and validation. Our results show an accuracy improvement of 37.1%, which is much better than that of ITU-T P.1203. This demonstrates the efficacy of the proposed method in overcoming the existing limitations. This is in line with the broader discourse on QoE estimation, which incorporates estimation parameters, as in Garcia et al. [8]. They explored the impact of the initial loading quality, congestion, and progressive video bitrate.

The similarities between the work of Raake et al. [7] and our own lie in our similar aim, similar acknowledgment of traditional method limitations, and emphasis on the importance of QoE prediction accuracy. By contrast, our work has a different focus, on leveraging facial emotion recognition to comprehend user experience and estimate QoE. Raake et al. [7] focused on a bitstream-based approach to estimate QoE extracted directly from video data, which is a purely technical approach. Raake et al. [7] used pre-existing datasets and avoided user involvement; however, we utilized real-world data collected under network condition observations and viewer facial emotion conditions. Next, our work leveraged several machine learning models utilizing user star ratings, information ad insertion, user feedback, and facial emotion recognition as influential features, in contrast to Raake et al. [7] who proposed a standardized statistical model based on technical video characteristics.

This study also contributes to the exploration of dynamic adaptive streaming, as discussed by Pereira and Pereira [9]. 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 using crowdsourcing. Our research uses adaptive streaming because it significantly improves QoE, as Oyman and Singh explained in their paper [12]. Yao et al. [13] explained the importance of using real-world bandwidth traces to obtain accurate HAS performance results. In our study, we used real-world bandwidth traces, in line with their testing. We used stats-for-nerds monitoring to evaluate and estimate video QoE for viewers.

In addition, our proposed method is in line with that employed in Ghani and Ajrash's [14] study on QoE prediction using alternative methods that do not rely on decrypted traffic data, such as psychophysiological measures, facial emotion recognition, and subjective feedback

from viewers under E2E environmental conditions. Porcu et al. [15] conducted a study to test this hypothesis. We believe that the QoE of end users may be made predictable by observing changes in facial emotions and using 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 positive Quality of Experience (QoE) [12,16]. Traditional methods that rely on network data go blind in this encrypted landscape, making it impossible to identify factors, such as intrusive ads, that negatively impact service users [8,10]. Thus, understanding user emotions is very important, as expressed by Zinner et al. [17].

With many users jockeying for bandwidth, competition to use bandwidth arises and can result in unfair allocation, resulting in QoE not being optimal for all users [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 significant factors that influence QoE.

Mapping audience expressions to measure QoE accurately requires a robust framework. Machine-learning algorithms were used in this study. Cohen's Fast Effective Rule Induction can be used as a powerful tool for extracting meaningful patterns from facial data, allowing us to map emotions to QoE levels [18]. The agreement between observers must be rigorously assessed to ensure the reliability of mapping. The statistics of Kappa Landis and Koch offer a well-established method for measuring the consistency of human judgments, which is essential for validating the accuracy of our proposed emotion-to-QoE approach [19].

Additionally, aligning our QoE assessments with established standards is critical for wider adoption. Bermudez et al. successfully applied the ITU-T P.1203 QoE model to live video streaming over LTE networks [20]. By adapting and integrating this standard framework into our FER-based approach, we recommend the development of compatibility with existing QoE measurement systems, paving the way for seamless integration into video-streaming platforms.

The term QoS is used to indicate the level of user satisfaction. Currently, the QoS metric indicates the level of user satisfaction by considering the user's background, psychological and emotional states, and expectancy. The QoS usually depends on several objective metrics, including delay variance or jitter, throughput, and delay. QoE is mostly subjective, indicating that the optimal QoS metric does not correspond to an enhanced QoE. QoE measures the level of user satisfaction with a certain service, which can be determined based on a subjective measurement, usually called the Mean Opinion Score (MOS), in which a user is asked to rank a video session between 1 (bad) and 5 (excellent). According to an MUX report on video streaming [9], a long rebuffering time is one of the main reasons why a user stops watching video content. Rebuffering leads to poor image quality and repeated playback errors. Sometimes, users stop watching because too many advertisements come onto their screen. Many solutions have been proposed to address these problems; however, the level of user engagement or satisfaction remains low because these solutions cannot perfectly handle network fluctuations and advertisement problems in real-time. Herein, we propose breakthroughs in ML that can handle the aforementioned issues by providing predicted insight into what QoE is experienced by users. Thus, video QoS and advertisement placement scenarios can be optimized to satisfy user QoE and automatically improve overall QoE. A list of content and advert combinations is presented in Table 1.

**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

Table 1. Cont.

Content Title	Content Length (s)	Number of Ad	Length of Ad	Position of Ad
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” [21]. 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 provides network operators with a sense of the network’s contribution to total customer satisfaction in terms of dependability, availability, scalability, speed, accuracy, and efficiency. Thus, many network researchers are working on this topic and attempting to incorporate it into network choices to ensure high customer satisfaction while using minimal resources.

Mean opinion 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) was proposed as a new measure of QoE with a new value of 0. According to the OS scale, the quality was 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 the user experience (UE) because numerous QoS factors directly or indirectly affect the user-perceived QoS. The key QoS parameters that impact multimedia services are bandwidth, jitter, delay, and packet loss rate.

Table 2 shows that the foundation of our proposal was fortified by insights from many studies in related areas. Amour et al. [22] introduced an improved QoE estimation method based on QoS and affective computing, emphasizing the relevance of emotional factors in the user experience. Bhattacharya et al. [23] highlighted 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. [15,24] and Antons et al. [25] provided notable points of view for estimating QoE using facial emotion gestures and electroencephalography, respectively, thereby reinforcing the importance of multimodal approaches.

Table 2. Our contribution among other related works.

Reference	Influence Factors	Considered Features
Amour et al. [22]	Resolution, bandwidth, delay	Face emotion
Bhattacharya et al. [23]	Delay, packet loss, bandwidth	Acoustic feature
Porcu et al. [15]	Delay, stalling	Face and gaze tracking information
Porcu et al. [24]	Blurring	Face and gaze tracking information
Antons et al. [25]	Noise signal	EEG
Kroupi et al. [26]	High quantization attribute	EEG
Arndt et al. [27]	Low bitrate encoding	EEG and EOG

Table 2. Cont.

Reference	Influence Factors	Considered Features
Arndt et al. [28]	Low bitrate encoding	EEG and gaze movement information
Engelke et al. [29]	Packet loss	Gaze movement information
Rai et al. [30]	High quantization attribute	Gaze movement information
Rai et al. [31]	Packet loss	Gaze movement information
Bailenson et al. [32]	Provoked delight and anxiety	Face emotion and 15 physiological features
Our Proposed Work	Video advertisement	Face emotion, video metadata and advertisement information

Moreover, the utilization of EEG correlates with video quality perception, as reported by Kroupi et al. [26]. Moreover, combining eye tracking with correlates brain activity to predict quality scores, as elaborated by Arndt et al. [27,28], underscores the interdisciplinary nature of our proposed method. Additionally, studies on the role of spatio-temporal distortions, as explained by Engelke et al. [29], and gaze disruptions in non-uniformly coded natural scenes [30,31] mirror our aim of 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 [32] contributes to the broader context of emotion-aware computing, supporting our approach of leveraging emotional cues for QoE prediction.

## 2.2. ITU-T P.1203 Standard

We compared 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 contains 4 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 [9].

The ITU-T P.1203 set of standards [33] developed by ITU-T is an example of a bitstream-based model. P.1203 is a quality model for HTTP-based adaptive audiovisual streaming [7,33,34]. 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 operating modes ranging from mode 0 to mode 3. The modes are defined by the amount of bitstream information provided, ranging from only the metadata (codec, resolution, bitrate frame rate, and segment time) in mode 0 to complete bitstream access in mode 3.

Herein, we focused only on the mode 0 Pv model. Mode 0 only requires metadata and is the quickest of all modes. However, the accuracy of mode 0 was shown to be 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 shows low accuracy in face emotion recognition (FER) and MOS. Hence, this study aims to improve accuracy by considering the results of the ITU-T P.1203 standard, the effect of the advertisement as another QoE influencing factor (IF), and the FER results.

### 2.3. Face Emotion Recognition (FER)

Facial expression recognition is one of the most important areas of research in human–computer interaction and human emotion detection [35]. A system must process various variations in the human face to detect facial expressions, such as color, texture, posture, expression, and orientation. To determine a person’s facial expression, 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 facial 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 subjective QoE assessment, which requires the user to fill out several questionnaires and provide ratings on a 0- to 5-star scale.

### 2.4. HTTP Adaptive Streaming (HAS)

YouTube is one of the most popular examples of an HAS application. YouTube has always used server-based streaming, but it has recently added HAS [36] as its default delivery/payout 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 owing to empty payout 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 fluctuating network circumstances. On the server, the video is divided into separate segments, each of which is available at different quality levels and representations (representing different bitrate levels). Based on network measurements, the client-side adaptation algorithm requests the next segment of the video at a bit rate level appropriate to the current network conditions [37].

### 2.5. QoE Influence Factors

Quality of experience (QoE) is important in determining user satisfaction with advertisements. This is an important indicator of the psychological expectations of a user’s fulfillment. To meet user expectations for high QoE, we can elaborate on several metrics [38,39], such as human IFs, which have the most complex parameters, such as system IFs. Human IFs include all the user information. Contextual IFs include information on location, user premise (watching environment), time spent watching (day or night), type of usage (casual watching, newly released favorite online gaming video), and consumption time (offload time and peak hours). System IFs include technical factors related to video quality that can be quantitatively measured using QoS measurements such as delay, jitter, packet loss, and throughput. Content IFs comprise the characteristics and information about the content of the video that users watch.

### 2.6. QoE Metrics

Several standards have been used in ITU-T and ITU-R for subjective video QoE testing, audio-visual video streaming, and subjective QoE grading techniques. Accordingly, grading techniques and testing standards have been proposed to enhance viewer satisfaction and MOS. Several factors may affect 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 primarily considers how packet or video traffic chunks travel in the network from the source to the destination. Alternatively, technical details include packet loss, jitter, delay, and throughput;
- MOS or subjective QoE measurement 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, examining the stalling frequency and stalling period.

### 2.7. QoE Assessment Types

As previously mentioned, QoE assessment techniques can be divided into subjective, objective, and hybrid assessments. Subjective measurements are time-consuming but can directly measure user satisfaction levels. They usually consider the user's preferences, age, individual psychology, viewing history, etc. 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 each 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 were performed using a computational model to estimate subjective test results. They aimed to predict the MOS that was as similar as possible to the real MOS obtained via subjective QoE measurements. The root-mean-square error (RMSE) [40] 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 [41]. Audio quality measurements include parametric, non-intrusive, and intrusive techniques. In intrusive techniques, the original signal is compared with the degraded signal in a test case. This technique yields accurate results but cannot be performed in real time. The non-intrusive technique estimates 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 [40].

Hybrid assessments offer advantages of subjective and objective practicality, portability, and convenience. Garcia et al. [8] defined an optimal approach for the QoE assessment of multiple-description coding [42,43] 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 method because it can accurately and directly collect information on 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 amounts of data and human resources are required, and this method cannot be applied in real-time. Owing to these limitations, the subjective assessment method is not widely used, except for verification and comparison with other methods.

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

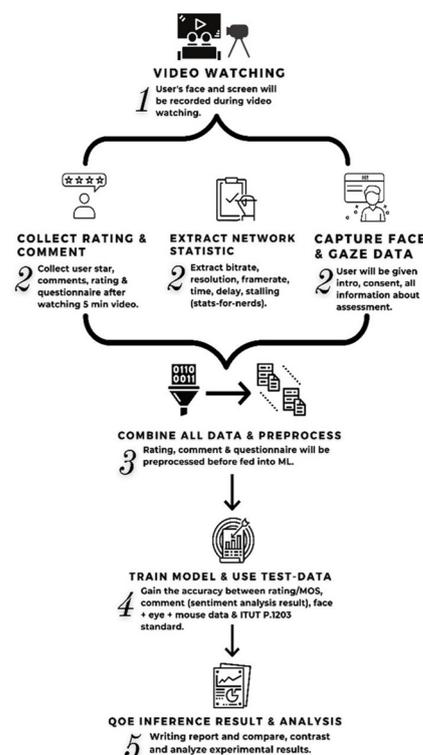
By considering user behavior when watching a video, the hybrid assessment method can be implemented in real time with high accuracy owing to supporting factors, such as artificial intelligence and statistics, which can reduce the limitations of both subjective and objective assessment methods. However, this method requires considerable amounts 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 to measure QoE in video streaming systems. To comply with the ITU standard, we used MOS to compare with the five-scale absolute category rating (ACR), as shown in Table 3. It is an ITU five-point quality scale [9].

**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 proposed method is illustrated in Figure 1. The first step consists of five steps. The steps are as follow: first, video watching session; second, data collection and storage; third, combining all data approaches; fourth, data pre-processing, data cleaning, training, and model evaluation; and finally, QoE estimation results and analysis.

**Figure 1.** Machine learning process summary.

These steps are elaborated in detail in Sections 3.1–3.5. To answer research question two (How can high-quality data be collected and prepared to yield an effective ML model for QoE estimation?) mentioned in the Introduction, we collected data from many sources to obtain many possibilities and answers to better train our model. A total of 700 participants from 114 countries completed the questionnaires, of which 125 were valid. Face and screen recordings were obtained from 60 participants, and 40 pairs of face and screen recordings were obtained.

Some data, from both video recordings and questionnaires, were eliminated due to data cleaning, pre-processing and separation from invalid data, blurry/low-brightness videos, and incompatible results. The methodology used during the initial stages of this study is illustrated in Figure 1; as shown in Figure 2, the user enters the laboratory room and reads the agreement. If they accept it, they proceed. They can leave it anytime and for any reason, if they do not accept it. We recruited participants by advertising the questionnaire and the survey to be conducted to individuals wanting to earn money (approximately USD 0.5).

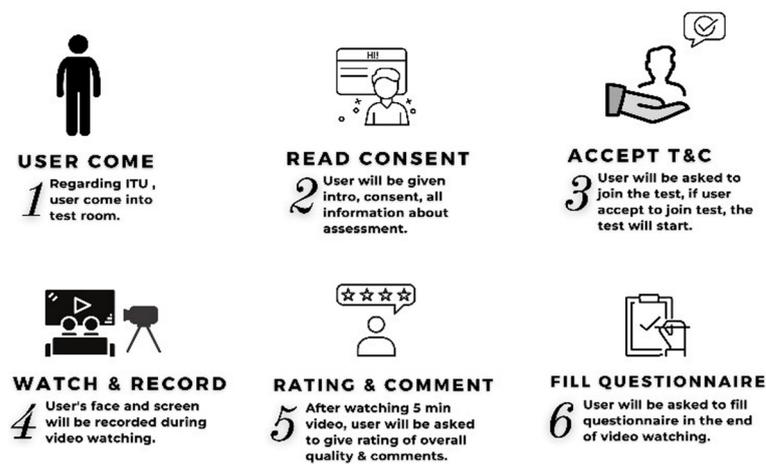


Figure 2. Data collection process.

All platforms are centralized at [www.tisaselma.online/form.html](http://www.tisaselma.online/form.html) (accessed on 20 February 2022) for Indonesians who speak Bahasa Indonesia or [www.tisaselma.online/form2.html](http://www.tisaselma.online/form2.html) (accessed on 20 February 2022) for English users. The automatic platform [www.xavyr.co.uk](http://www.xavyr.co.uk) (closed for public use) (accessed on 20 February 2022), was developed for automatic facial and screen recordings. The details of the content and advertisements are summarized in Table 1.

Moreover, if the participants provided consent, we recorded their faces and screens 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 to be fed into the ITU-T P.1203 model.

All 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 differed. The video content titles and advertisement data details are presented in Section 4 and Table 9, respectively. We chose as a disturbance factor an advertisement that does not relate to the content, has many repetitions, is located in the beginning, middle, or end of the content, and is free to use and can be downloaded from the Internet. All participants were asked to watch five videos for approximately five minutes without advertisements. The advertisements were displayed for approximately 7–11 min.

### 3.1. Video Watching Session

If the participants agreed to the terms and conditions, they watched the video while we collected video recordings of their face and the screen. This process is summarized in Figure 2.

### 3.2. Data Collection and Storing

All star ratings and comments were collected from the participants after watching each video. Appendix B presents some of the questions. The complete results and analyses will be elaborated upon in another journal.

### 3.3. Combine All Data Approach

From the video screen recording, we can extract stats-for-nerds to obtain bitrate, resolution, and frame rate for every 4 s window. The 4 s windows were utilized as a period to ease the prediction process by down-sampling. To extract the video statistics, we captured 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 act as the input

for the ITU-T P.1203 model, we can right-click on the YouTube player and turn on the stats-for-nerds option, as shown in Figure 3. The statistics for stats-for-nerds are presented in Figure 4. The results of the extraction of stats-for-nerds input and output are shown in Figure 5. We captured network condition statistics for every second during playback in the video watching period.

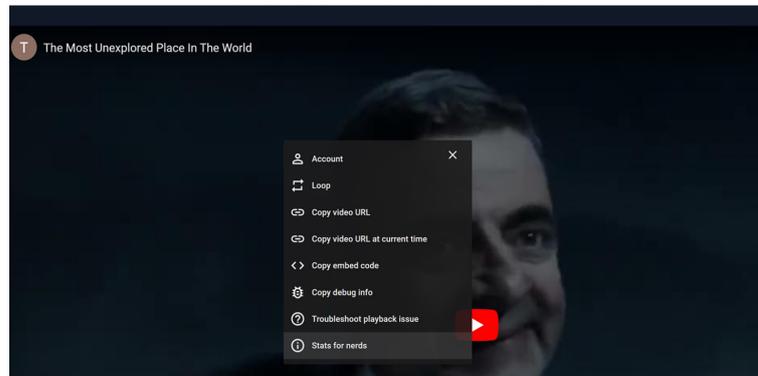


Figure 3. Stats-for-nerds on YouTube player, summoned by right-clicking the player.

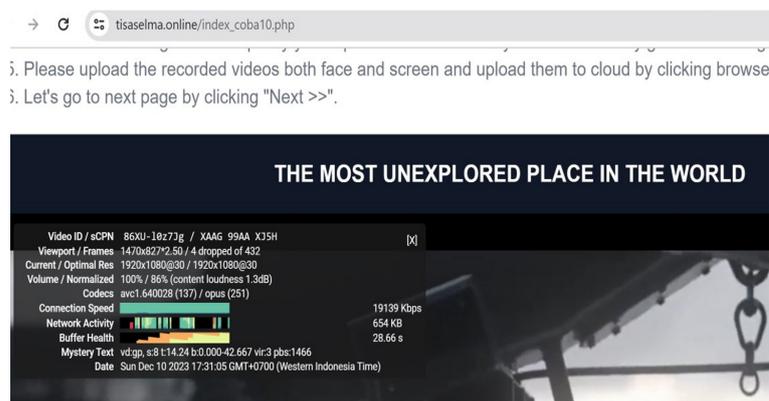


Figure 4. Stats-for-nerds appears in the right corner of the YouTube player.

```

1  {
2    "I11": {
3      "segments": [
4      ],
5      "streamId": 42
6    },
7  },
8  "I13": {
9    "segments": [
10   {
11     "bitrate": 5000,
12     "codec": "h264",
13     "duration": 4.0,
14     "fps": 30.0,
15     "resolution": "1280x720",
16     "start": 0
17   },
18   {
19     "bitrate": 5000,
20     "codec": "h264",
21     "duration": 4.0,
22     "fps": 30.0,
23     "resolution": "1280x720",
337   5.0,
338   5.0,
339   5.0,
340   5.0,
341   5.0,
342   5.0,
343   5.0,
344   5.0,
345   5.0,
346   5.0
347   ],
348   "O35": 5.000000000000002,
349   "O46": 4.836041457562939,
350   "date": "2022-04-23T17:01:05.482085",
351   "mode": 0,
352   "streamId": 42
353   },
354   "tisa6.mp4": {
355     "O23": 5.0,
356     "O34": [
357       5.0,
358       5.0,
359       5.0,

```

Figure 5. Screenshot of input and output \*.json file from ITU-T P.1203. On the left is the input, and on the right is the output file.

When the user was watching, we captured video statistics and simultaneously recorded the user's face. We used VLC to extract frames from the video. We then obtained information that needed to be fed into the ML model: bitrate, frame rate, and resolution. The resulting features that were extracted from stats-for-nerds for each video from all participants are listed in Table 4. Next, we fed video statistics, including all video metadata (\*.json file), into the ITU-T 1203 machine learning set up to predict Mean Opinion Score (MOS). The ITU-T 1203 machine learning model no longer requires training during this step.

**Table 4.** Sample features extracted from stats-for-nerds for each video and for 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

The main competing approach used the ITU P.1203 standard. The input features for ITU are a \*.json file containing frame rate, bitrate, resolution, stall length, and stall position. The input file is shown in Figure 5; we can see that for every 4 s window, the frame rate, bitrate, resolution, stalling length, and position may fluctuate. The input of each 4-s window is obtained from YouTube stats-for-nerds, and will be given a 1–5 ACR score by the ITU-T P.1203 model. The following standards were implemented using the assessment software [3,44]:

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 a dependable transport video quality estimation module;
3. ITU-T Rec. P.1203.2—Audio quality estimate module for metric bitstream-based quality evaluation of progressive download and adaptive audio–visual 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 audio–visual streaming services over dependable transport.

### 3.4. Pre-Processing, Data Cleaning, Model Training and Evaluation

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

For the pre-processing step, we combined all features of video statistics, star ratings, and emotion recognition tests to compare them using the same metric: ACR score. After extracting frames from the face and video recordings using VLC, we input all resulting frames into the DeepFace machine learning model and performed several training and testing steps. We then selected the 8 best attributes using the symmetrical t attribute value, ranked using 10-fold cross-validation; the best 9 that were ranked using Relief F Attribute Eval, Ranker via 10-fold cross-validation; and the best 13 that were yielded by correlation attribute val using 10-fold cross-validation. These attributes are: long 5 min ad, FER, ad loc, name, bitrate, length ad, resolution, title, ad count, content length, ITU-res, ad each min, and repeat.

The frames were extracted from the face videos and fed into the CNN/DeepFace model for FER. In this step, the model was trained to satisfy certain conditions. The outputs included seven emotions: happiness, surprise, disgust, sadness, fear, neutrality, and anger. We mapped all emotions to MOS based on their relatedness. The ACR scores are shown in Table 2, and the mapping from the DeepFace emotion to the ACR score is shown 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

However, due to time and cost limitations, we will attempt to improve the dataset in the future. It was not easy to obtain consent from many participants to use their face recordings for our research. It took approximately half a year to one year to gather more than 600 participants to complete our questionnaire, and 50 participants to provide their face and screen recordings. Some participants came to our laboratory, while others conducted the test online using a unified, self-made platform.

However, collecting the required data posed a significant challenge. Our dataset is unique, and we have not identified any existing dataset that aligns directly with the specifics of our research. Considering time and budget constraints, augmenting our dataset remains a focal point for future research. Securing consent from a significant number of participants to use their face recordings is demanding, typically requiring six months to a year to gather over 600 questionnaire responses and obtain face recordings from 50 participants. For these 50 participants, we intended for them to watch five videos of 10 min each. The result was thus 50 multiplied by five (for the videos), resulting in 250 videos. Each video was approximately 10 min long before cleaning and pre-processing. The raw videos contained almost two hours of facial recording sessions. Recordings of screens while watching the videos were excluded. We derived a total of approximately 500 videos. Screen recording videos were obtained to determine attention and perspective while using our platform. We will retain these data for user engagement optimization in the future.

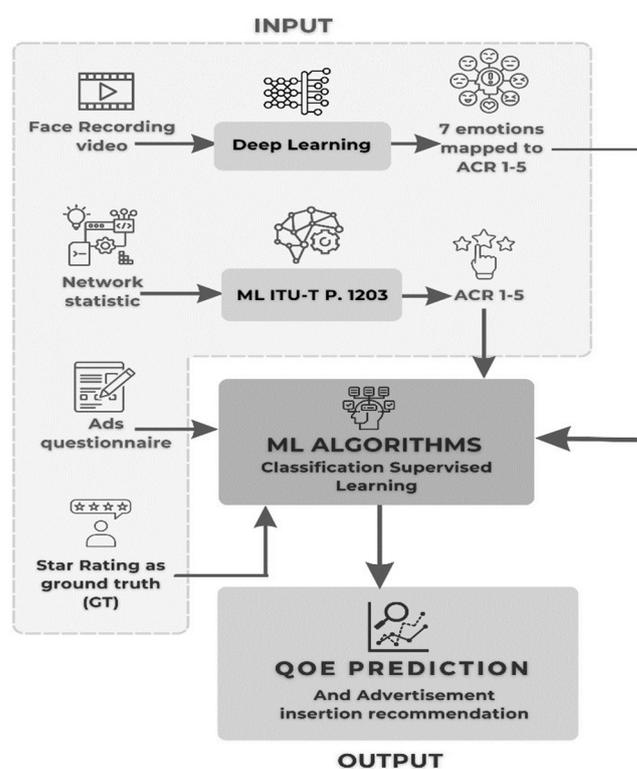
In the future, we will attempt to expand and vary our video dataset to obtain better accuracy and more robust facial emotion recognition models. First, we generated 1 million synthetic data points using a generative AI model; however, we needed more time to determine the best hyperparameter to tune with our model. Hence, we plan to publish the results of future studies. Next, we plan to perform emotion simulations by utilizing such tools and techniques as facial action coding to generate synthetic facial expressions representing diverse emotions, and data augmentation to artificially increase the size and diversity of our dataset without collecting new data. Additionally, we plan to implement a well-structured data organization system to enable efficient annotation, analysis, and model training. Moreover, we have balanced our data by applying a cost-sensitive support vector machine (CS-SVM).

### 3.5. QoE Evaluation Result and Analysis

The six emotions listed in DeepFace were mapped to ACR scores, as listed in Table 5. From DeepFace, we obtained MOSs of 1-5. These networks extract features from questionnaires, such as demographics and user preferences for ads, videos, placements, and advertisement matches with video content. User preferences can reveal user expectations and perceptions related to streaming video services' quality. The outputs from these three networks were then fed into the QoE prediction network. The QoE prediction network was used to infer QoE scores in relation to advertisements, video content, network conditions, and user emotional states. The QoE prediction networks were trained using QoE scores

collected from star ratings provided by users after watching each video. We evaluated our proposed architecture using user-generated facial recordings, video advertisements, questionnaires, and QoE star rating scores. We compared our architectures with traditional QoE prediction methods, and showed that our architecture achieved much better results as a base reference. In addition, our architecture can capture complex relationship between QoE and its influencing factors, such as facial expressions, video advertisement content, and user preferences. This ability to capture complex relationships is related to the use of deep learning and machine learning algorithms, which can learn nonlinear relationships from raw data.

Moreover, as mentioned in the architecture, we used DeepFace to predict user emotions automatically. DeepFace is a deep learning-based facial recognition system developed in 2014 by Facebook Research with a high accuracy of up to 97.35% on Labeled Faces in the Wild (LFW), which performance is better than humans can achieve. The DeepFace architecture consists of several convolutional and pooling layers, three locally connected layers, and a final Softmax layer. The input to the network is a face image of  $152 \times 152$  RGB. The proposed architecture is illustrated in Figure 6 below.



**Figure 6.** Model architecture.

First, the network extracts low-level features from images using convolutional layers. Second, these features are downsampled using pooling layers. Locally connected layers are used to extract higher-level features from low-level features. Third, a softmax layer is used to classify the input image as being of faces/emotions in the training data. DeepFace was trained using stochastic gradient descent on a dataset of over 4 million faces. This dataset included faces differing in age, ethnicity, and sex.

DeepFace was implemented using the Caffe deep-learning framework, and linear units (ReLUs) were rectified as activation functions. It pools layers with max pooling with a stride of two locally connected layers using a  $3 \times 3$  kernel. The Softmax layer had 128 outputs related to the number of faces in the training data.

The overall system architecture is illustrated in Figure 6. Machine learning was applied at two locations in the system. First, as shown in Figure 6, a deep learning model was

used to obtain emotional features. In this case, the deep learning model was pretrained; therefore, we did not use it to learn from new data. It was used to extract features from videos. Second, we used machine learning to predict QoE. To achieve acceptable prediction quality, we trained a machine learning model using the different types of collected data, and carefully selected the model parameters.

In this study, 18 algorithms are compared. We found that the Random Forest classifier outperformed other algorithms. Random forests handle nonlinear relationships and interactions between diverse inputs. Random Forests work by constructing many decision trees during the training and outputting of the class, which are the classes of individual trees. By aggregating the predictions of many trees, the variance between predictions is reduced, and the accuracy can be improved over a single decision tree.

Random Forest is an ensemble learning method that is well suited to handling complex and multidimensional datasets, making it a strong candidate for predicting video Quality of Experience (QoE) with diverse inputs, as presented in this case. Facial emotions directly measure users' emotional engagement with video content. Network statistics and ITU-T P.1203 results highlight the technical quality of streaming services. Advertisement questionnaire and "stats-for-nerds" data offer insights into user preferences and their interactions with content. The star rating grounds the model in user feedback, ensuring that the predictions have a tangible reference point. The classifier takes all of these inputs, learns from the complex relationships among them, and outputs a prediction that encapsulates the entire user experience.

The process by which Random Forest handles inputs and produces outputs can be divided into three steps. First, data wrangling comprises feature engineering, missing value handling, and data splitting. Second, random forest construction consists of building trees, growing trees, and ensemble power. The last can be summarized as comprising three steps: new video encounters, collective wisdom, and model performance. The process was conducted as follows.

In the feature-engineering process, each attribute requires a specific treatment. Numerical features, such as bitrate and resolution, can be normalized for scale, whereas categorical features, such as emotions, can be one-hot encoded. The star ratings representing the target class (QoE) remain untouched. In missing-value handling, imputation techniques such as mean/median filling or more sophisticated methods such as KNN imputation can be employed if any data point is missing. Next, we split 215 rows into training and testing sets. The training set (we used trial and error to find the best split, including 80:20 Pareto, 94:6 ratio, 92%:8% split, and 50- and 60-fold cross-validation) was used to train the model. In contrast, the testing set (20–30%) was used for an unbiased evaluation of its performance.

Moreover, in random forest construction, especially in building tree steps, the algorithm randomly selects subsets of features (typically, the square root of the total number). It builds multiple decision trees (forest). Each tree uses these features to independently create a set of "rules" to predict QoE. These rules might split the data based on, for example, a bitrate exceeding a threshold or a specific emotion being present. The next step was to grow trees. In this step, each tree grows until it reaches a predefined maximum depth or a stopping criterion such as purity (all data points in a leaf node belong to the same class). Moreover, hundreds or even thousands of trees are created during the ensemble power step, forming a random forest.

In the prediction and evaluation step, a new video encounter phase starts when a new video with its corresponding 17 attributes is presented; each tree in the forest makes its own QoE prediction based on its learned "rules". In the collective wisdom step, the final predicted QoE is the average of all the individual tree predictions, leveraging the collective wisdom of the forest to obtain a robust and stable outcome. In the model accuracy measurement step, the predicted QoE derived from the testing set is compared with the actual star ratings using metrics such as Root Mean Squared Error (MSE), precision, and recall. These metrics assess how closely the model's predictions align with real user perceived QoE. We chose Random Forests, among others, to achieve the highest accuracy,

reveal which features are most influential in predicting QoE, and provide valuable insights into user experience and potential areas for future improvement. In our case, the most influential features based on the random forest algorithm were `participant_id` and `video_id`, because we numerically encoded all participants and videos. This may mean that the preference for giving a 1–5 rating is very subjective based on video content.

### 3.6. Analysis of Machine Learning Methodologies, Features Importance, and QoE Perceptions

To improve end-user video Quality of Experience (QoE) perception, leveraging machine learning explanations and feature importance analysis can provide valuable insights. Here, a more in-depth analysis of how these aspects can achieve better QoE perception using facial emotion recognition, advertisement insertion information, and network conditions is presented.

In face emotion recognition, feature importance is used to identify key facial features, consisting of understanding which facial features have the highest impact on emotion recognition, which can enhance the interpretability of our model. By building our proposed approach, we can determine how facial expressions affect perceived content quality. Our future work will focus on real-time adjustments by developing a system that tracks dynamic changes in facial expressions. This enables real-time adjustments to content delivery, such as aligning content to match the emotional state of the end user.

Our proposed approach leverages multimodal analysis by integrating facial emotion recognition, ad insertion details, and network conditions into a holistic QoE model to obtain a more comprehensive view of the end-user experience. We have identified relevant features by analyzing which contribute the most to determining ad relevance, such as context, viewer demographics, or emotional state, which can help optimize ad insertion strategies. By analyzing the questionnaire results, we found that most users hated unskippable ads, at approximately 32%. Next, we can infer user perceptions of QoE by considering the network conditions pertaining while the user is watching one video session. We investigated the relationship between content delivery strategies and ad-insertion scenarios using user ratings. We found that users could still be happy while watching ads for up to 10 s, at approximately 43%.

We used neutral content and advertisements to clarify the facial emotions triggered by network conditions or advertisement scenarios. Emotions related to network conditions may be triggered negatively if the network (bandwidth, resolution, latency, stalling time, stalling frequency, etc.) worsens. We can see that many advertisements may influence emotions related to advertisement insertion in mid-roll, unskippable, or unrelated advertisements that are too long during watching sessions.

Feature importance and accuracy are crucial for 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 provided valuable insights. Han and Yu [45] provided a theoretical framework explaining the relationship between the stability and accuracy of feature selection, emphasizing the dependency of feature selection stability on the sample size. This highlights the importance of considering the stability of feature-selection methods in relation to accuracy, particularly in the context of varying sample sizes.

Strobl et al. [46] highlighted a bias in variable importance measures towards correlated predictor variables. This bias can affect the accuracy of the feature selection methods, indicating the need to account for correlations between predictor variables when assessing feature importance to ensure accurate and reliable results. Furthermore, Altmann et al. [47] 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. [48] compared Random Forest and its Gini importance with standard chemometric methods for feature selection and the classification of spectral data, indicating a preference for Gini feature importance as a ranking criterion because

it considers 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 and involves considerations such as stability, bias, correlation among predictor variables, and the impact of feature selection on classification accuracy. Understanding and addressing these factors is essential for optimizing feature selection methods and improving the accuracy of machine learning models.

#### 4. Experimental Results

##### 4.1. Survey Results and Statistics

For this investigation into advertising, questions were designed to ensure the accuracy of the data obtained. The results of this survey are summarized in Tables 6–8.

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

The Most Annoying Advertisement Type from 122 Participants		
Case	Participants	Percentage (%)
Many repeated advertisements at 1 point in time in mid-roll	22	18.03%
Single 5 min advertisement long in mid-roll	22	18.03%
In 5 min of video content, every 1 min, 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 were completed with video recordings

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

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

The information is only partly listed here because of page limitations. To show the complete data would require hundreds of rows. A condensed version of the comparison between the ITU-T P.1203 face emotion recognition results, star ratings as ground truth given by participants, and our predictions can be seen in Table A2 in Appendix C.

In Table 9, we can see the title, video resolution, video bitrate, ITU-T P.1203 result (ACR score), face emotion recognition result (ACR score), content length (in seconds), the number of advertisements in one video content, the length of advertisement (in seconds), advertisement location, repeated advertisement (1 is for repeating and 0 is no repetition of advertisement), the presence of a five-minute advertisement in one video (1 is for present, 0 is for absent), the presence of advertisements in each minute 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 absent), the presence of unskippable advertisements in one video (1 is for present, 0 is absent), and the test case or ground truth in supervised classification, with a star rating of 1–5. All these parameters are used in machine learning to predict the end-user QoE. In Table 10, we list all the features used in our approach.

**Table 9.** Sample list of features including ITU-T P.1203, video metadata, FER values and star rating.

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

**Table 10.** List of features fed into the ML model.

Amount	Attributes
16	user.id, video.id, resolution, bitrate, fer, content.length, ad.loc, ad.count, long.5min.ad, ad.each.min, pre.mid.post.same.id, repeated.ad, no.skip, stars
15	user.id, video.id, itu, resolution, bitrate, fer, content.length, ad.loc, ad.count, long.5min.ad, ad.each.min, pre.mid.post.same.id, repeated.ad, no.skip, stars
8 Selected Attributes Using Symmetrical Uncert Attribute Eval, Ranker, 10-fold cross-validation	video.id, fer, content.length, ad.loc, long.5min.ad, ad.each.min, pre.mid.post.same.id, repeated.ad, stars
9 Selected Attributes Using Relief F Attribute Eval, Ranker, 10-fold cross-validation	video.id, content.length, ad.loc, ad.count, long.ad, long.5min.ad, ad.each.min, resolution, bitrate, stars
13 Selected Attributes Using Correlation Attribute Eval Using Ranker, 10-fold cross-validation	user.id, video.id, itu, resolution, bitrate, fer, content.length, ad.loc, ad.count, long.5min.ad, ad.each.min, pre.mid.post.same.id, repeated.ad, stars

Table 11 summarizes all the results, and we have categorized the machine learning results via 15 attributes. The 15 attributes are listed in Table 10. We conducted several experiments using selection attribute algorithms, and found the three best algorithms that resulted in the best accuracy using the best number of attributes. Finally, we arrived at 8, 9, and 13 attributes that are better because the other attributes showed zero correlations, zero merit average, and bad ranks. The results of attribute selection are presented in Table 12. The best accuracy for each category is highlighted in gray.

**Table 11.** Summary of machine learning accuracy results using 15 attributes.

ML Method	Naïve Bayes Updateable			Multi-Layer Perceptron CS			Meta Random Subspace		
Test Types	94:6	60-Fold	Train Set	94:6	60-Fold	Train Set	94:6	60-Fold	Train Set
CCI <sup>1</sup>	69.23%	46.05%	59.53%	61.54%	51.63%	100.00%	61.54%	52.56%	60%
ICI <sup>2</sup>	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 <sup>3</sup>	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-fold	Train Set	94:6	60-fold	Train Set	94:6	60-fold	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-fold	Train Set	94:6	60-fold	Train Set	94:6	60-fold	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

Table 12. Summary of machine learning results with attribute selection classification.

8 Selected Attribute Using Symmetrical Uncert Attribute Eval, Ranker, 10-Fold Cross-Validation								
Test Types	Random Forest		Furia		Jrip		Meta Decorate	
	60-Fold	94:6	60-Fold	94:6	60-Fold	94:6	60-Fold	94:6
CCI	53.85%	33.02%	45.58%	53.85%	45.58%	53.85%	48.37%	53.85%
ICI	46.15%	66.98%	54.42%	46.15%	54.42%	46.15%	51.63%	46.15%
RMSE	0.3698	0.3865	0.4197	0.4136	0.37	0.3774	0.3666	0.3578
Total Instances	13	215	215	13	215	13	215	13
Precision	N/A	0.298	N/A	N/A	0.543	N/A	0.491	N/A
Recall	0.538	0.33	0.456	0.538	0.456	0.538	0.484	0.538
Test Types	SMO		Tree SPAARC		Tree Optimized Forest		Local KNN	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	48.84%	53.85%	42.33%	46.15%	34.88%	53.85%	16.74%	15.38%
ICI	51.16%	46.15%	57.67%	53.85%	65.12%	46.15%	83.26%	84.62%
RMSE	0.3705	0.3823	0.3749	0.3845	0.4157	0.3724	0.396	0.4038
Total Instances	215	13	215	13	215	13	215	13
Precision	0.406	N/A	N/A	N/A	0.328	N/A	0.302	N/A
Recall	0.488	0.538	0.423	0.462	0.349	0.538	0.167	0.154
Test Types	Multi-Layer Perceptron		Naïve Bayes		Chirp		Multi Class Classifier	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	38.60%	53.85%	43.26%	53.85%	31.63%	38.46%	44.65%	53.85%
ICI	61.40%	46.15%	56.74%	46.15%	68.37%	61.54%	55.35%	46.15%
RMSE	0.3867	0.3637	0.3872	0.373	0.523	0.4961	0.3762	0.378
Total Instances	215	13	215	13	215	13	215	13
Precision	0.375	N/A	0.416	N/A	0.32	N/A	0.428	N/A
Recall	0.386	0.538	0.433	0.538	0.316	0.358	0.447	0.538
9 Selected Attributes Using Relief F Attribute Eval, Ranker, 10-fold cross-validation								
Test Types	Random Forest		Furia		Jrip		Meta Decorate	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	37.21%	46.15%	53.85%	47.44%	47.44%	53.85%	46.98%	61.54%
ICI	62.79%	53.85%	46.15%	52.56%	52.56%	46.15%	53.02%	38.46%
RMSE	0.3909	0.3953	0.3778	0.3993	0.3666	0.3759	0.3741	0.3603
Total Instances	215	13	13	215	21	13	215	13
Precision	0.375	N/A	N/A	N/A	N/A	N/A	0.461	N/A
Recall	0.338	0.462	0.538	0.474	0.474	0.538	0.47	0.615
Test Types	SMO		Tree SPAARC		Tree Optimized Forest		Local KNN	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	48.84%	53.85%	41.40%	46.15%	35.81%	38.46%	46.05%	7.69%
ICI	51.16%	46.15%	58.60%	53.85%	64.19%	61.54%	53.95%	92.31%
RMSE	0.3683	0.3783	0.3756	0.3845	0.4162	0.4009	0.373	0.4175
Total Instances	215	13	215	13	215	13	215	13
Precision	0.414	N/A	N/A	N/A	0.335	N/A	0.379	N/A
Recall	0.488	0.538	0.414	0.462	0.338	0.385	0.46	0.077
Test Types	Multi-Layer Perceptron		Naïve Bayes		Chirp		Multi Class Classifier	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	42.33%	61.54%	39.07%	61.54%	34.42%	30.77%	46.05%	61.54%
ICI	57.67%	38.46%	60.93%	38.46%	65.58%	69.23%	53.95%	38.46%
RMSE	0.3807	0.3661	0.4029	0.3605	0.5122	0.5262	0.373	0.3706
Total Instances	215	13	215	13	215	13	215	13
Precision	0.379	N/A	0.398	N/A	0.393	N/A	0.379	N/A
Recall	0.423	0.615	0.391	0.615	0.344	0.308	0.46	0.615
13 Selected Attributes Using Correlation Attribute Eval Using Ranker, 10-fold cross-validation								
Test Types	Random Forest		Furia		Jrip		Meta Decorate	
	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	30.70%	46.15%	48.37%	53.85%	46.98%	53.85%	44.19%	53.85%
ICI	69.30%	53.85%	51.63%	46.15%	53.02%	46.15%	55.81%	46.15%
RMSE	0.4179	0.4119	0.4163	0.4033	0.3671	0.3759	0.3715	0.3598
Total Instances	215	13	215	13	215	13	215	13
Precision	0.297	N/A	N/A	N/A	N/A	N/A	0.408	N/A
Recall	0.307	0.462	0.484	0.538	0.47	0.538	0.442	0.538

Table 12. Cont.

8 Selected Attribute Using Symmetrical Uncert Attribute Eval, Ranker, 10-Fold Cross-Validation								
Test Types	Random Forest		Furia		Jrip		Meta Decorate	
	60-Fold	94:6	60-Fold	94:6	60-Fold	94:6	60-Fold	94:6
	SMO		Tree SPAARC		Tree Optimized Forest		Local KNN	
Test Types	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	47.44%	53.85%	44.65%	46.15%	31.63%	38.46%	34.88%	30.77%
ICI	52.56%	46.15%	55.35%	53.85%	68.37%	61.54%	65.12%	69.23%
RMSE	0.3689	0.3783	0.3708	0.3845	0.4293	0.4234	0.4335	0.4356
Total Instances	215	13	215	13	215	13	215	13
Precision	0.405	N/A	N/A	N/A	0.331	0.346	0.405	0.354
Recall	0.474	0.538	0.447	0.462	0.316	0.385	0.349	0.308
	Multi-Layer Perceptron		Naïve Bayes Simple		Chirp		Multi Class Classifier	
Test Types	60-fold	94:6	60-fold	94:6	60-fold	94:6	60-fold	94:6
CCI	33.95%	53.85%	38.14%	46.15%	36.74%	38.46%	44.19%	46.15%
ICI	66.05%	46.15%	61.86%	53.85%	63.26%	61.54%	55.81%	53.85%
RMSE	0.4224	0.3887	0.4093	0.3698	0.503	0.4961	0.3783	0.376
Total Instances	215	13	215	13	215	13	215	13
Precision	0.322	N/A	0.394	N/A	0.339	N/A	0.407	N/A
Recall	0.34	0.538	0.381	0.462	0.367	0.385	0.442	0.462

<sup>1</sup> CCI stands for correctly classified instances. <sup>2</sup> ICI stands for incorrectly classified instances. <sup>3</sup> N/A stands for Not Available.

#### 4.2. Competing Approaches

We used several machine-learning models to investigate their accuracy against our dataset, including TreeSPAARC. These included 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 on Windows 11 Pro, 21H2, Intel(R) Core(TM) i7-1065G7 CPU @ 1.30 GHz 1.50 GHz. Some notable parameters used in our approach 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

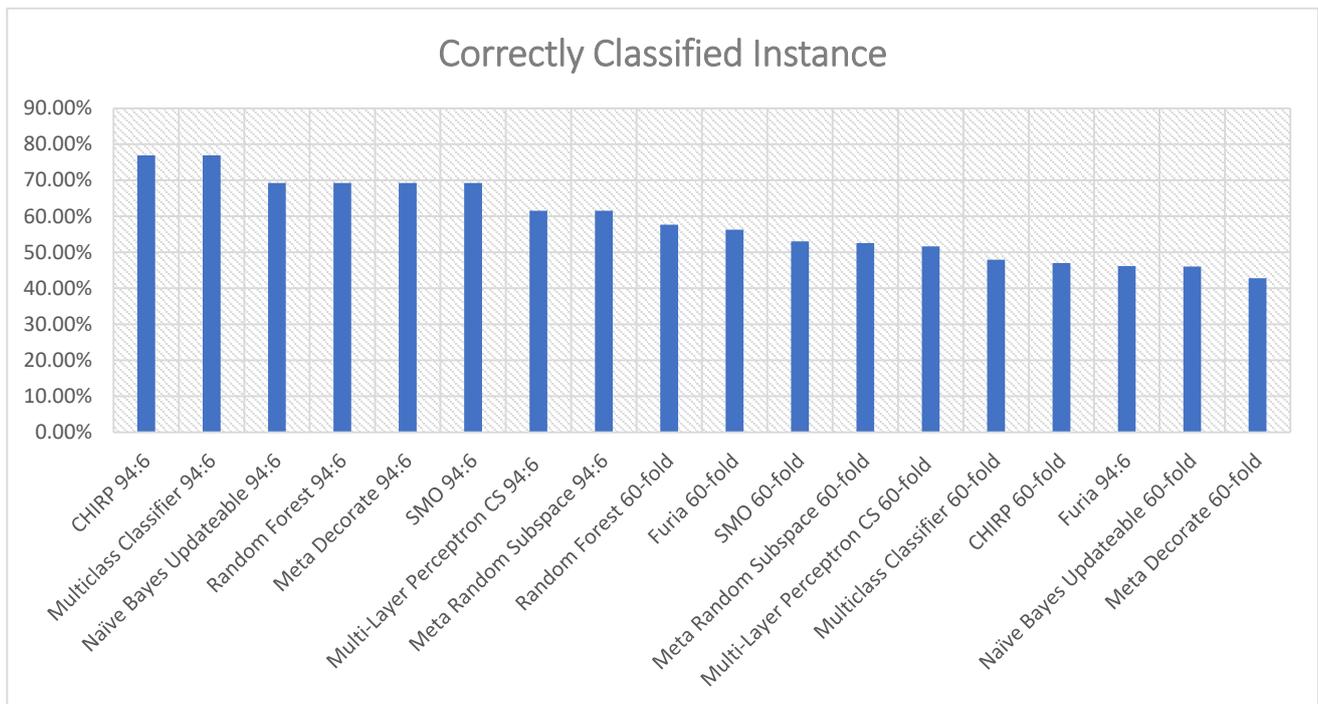
In this section, we elaborate on the experimental results. Our evaluation employed Pareto ratios of 80:20 and 94:6 to train and test the split dataset and perform 60 cross-validations. To evaluate the accuracy of our approach, we correctly and incorrectly classified instances via Root Mean Squared Error (RMSE), precision, and recall. The evaluation results are summarized in Tables 11 and 12.

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

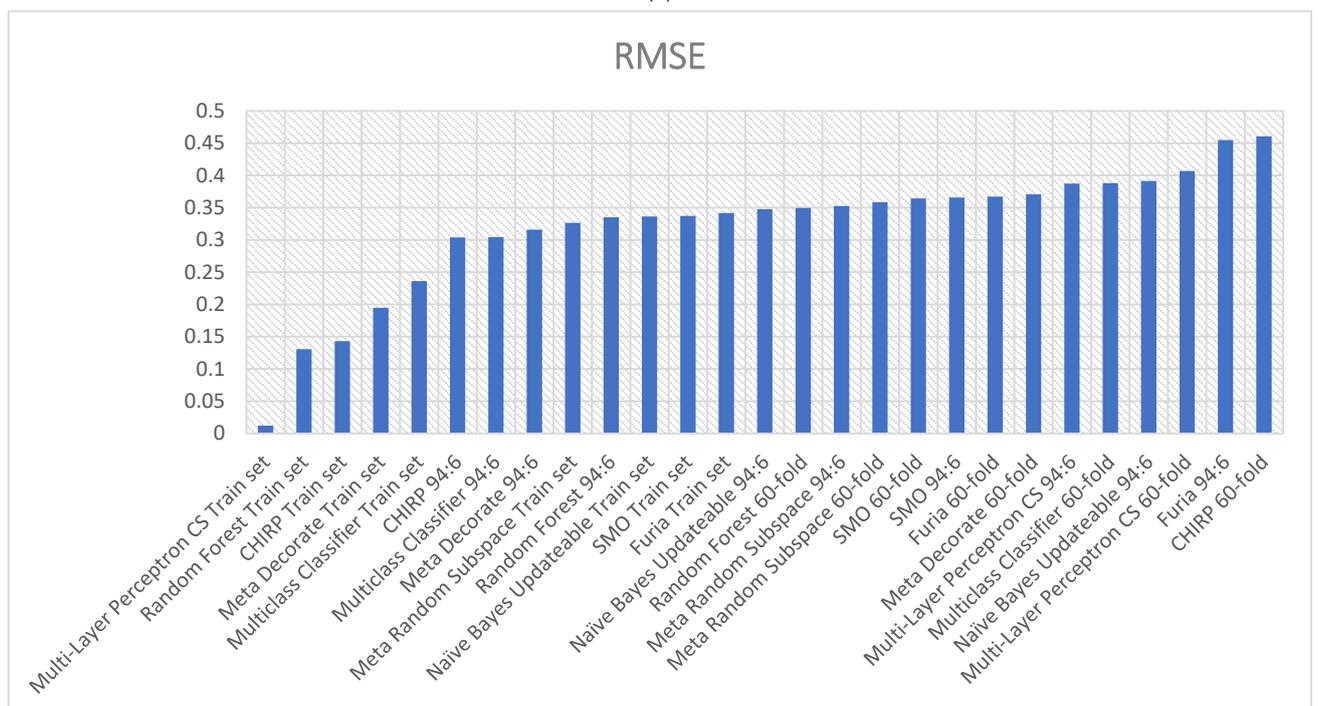
A better model has a lower error, such as that indicated by RMSE; the better the model, the lower the error, or RMSE value [33]. A graph of the correctly classified instances is shown in Figure 7a. A summary of the RMSE results is shown in Figure 7b. Figure 7c,d show summaries of precision and recall. Precision is the ratio of True Positives to all Positives. In our issue statement, the star ratings have been accurately identified as the

subjective MOS star ratings among all participants. Mathematically, the precision can be formulated as true positives divided by the summation of true positives and false positives.

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

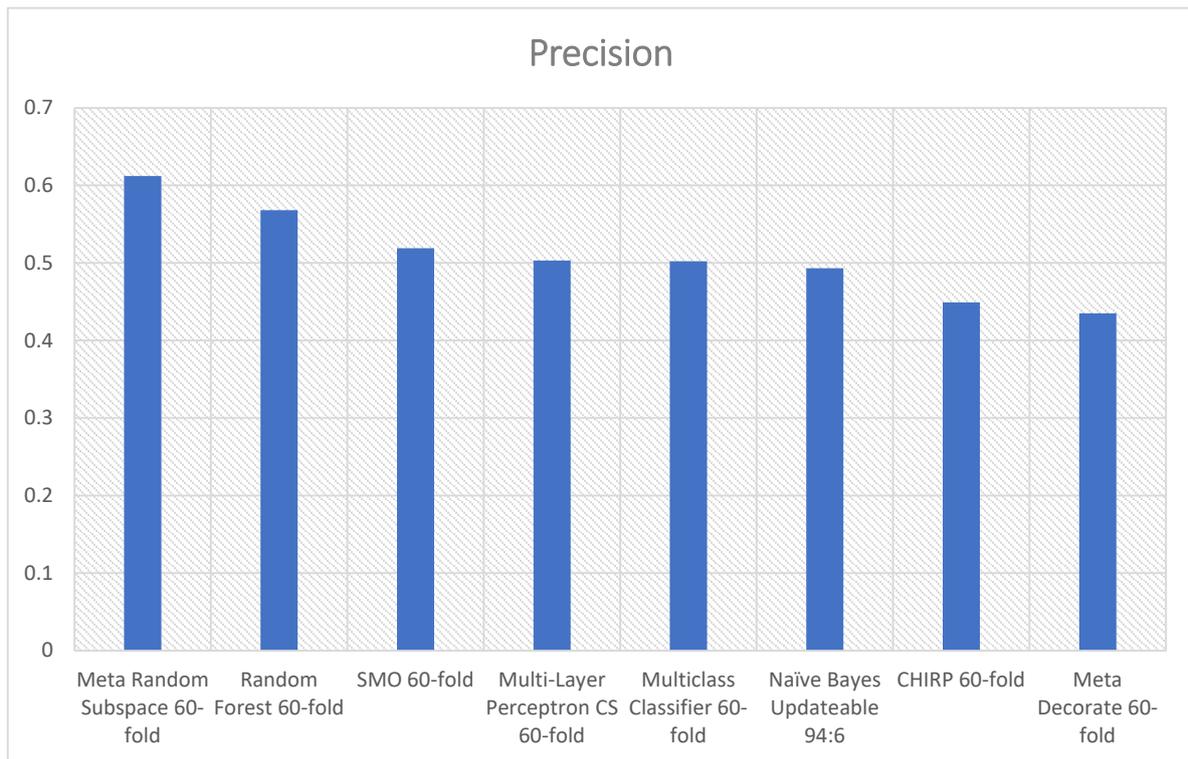


(a)

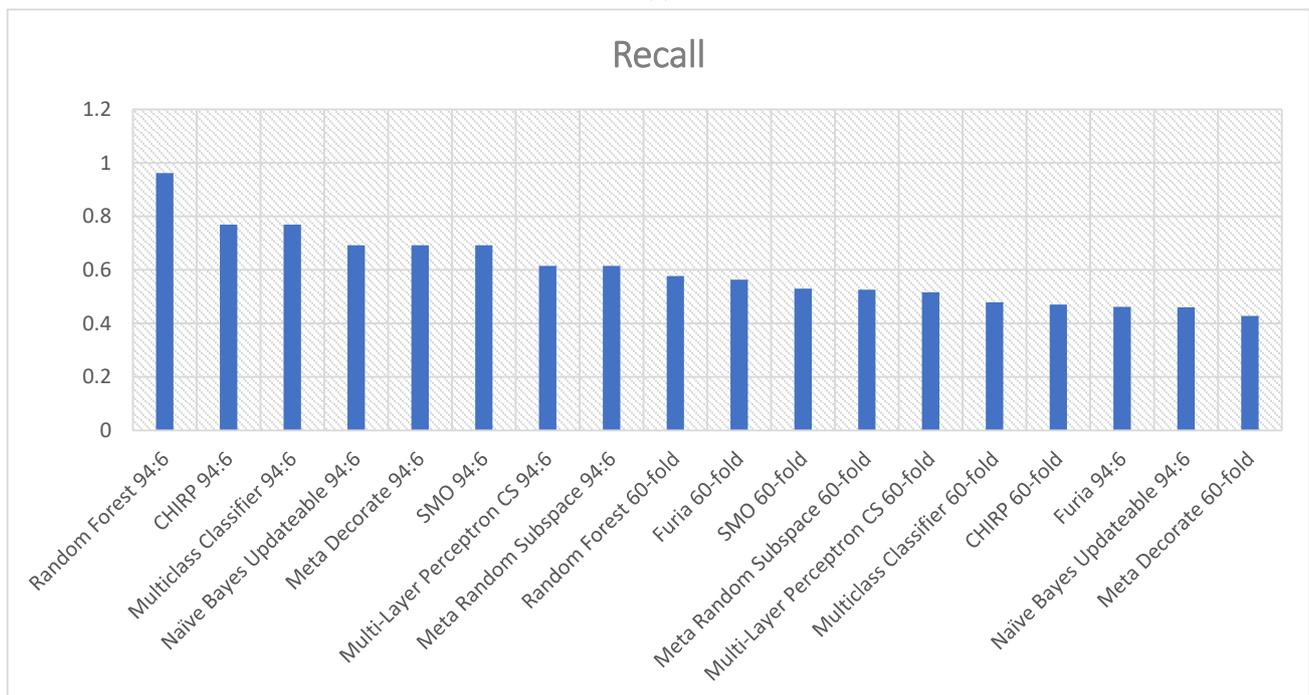


(b)

Figure 7. Cont.



(c)



(d)

**Figure 7.** Accuracy evaluation measurement for the 15 selected attributes using 15 state-of-the-art machine learning models. We chose to show only the bar chart utilizing 15 attributes due to it having the highest accuracy. The measurements we used include (a) correctly classified instances, (b) RMSE, (c) precision, and (d) recall.

On the other hand, recall measures how well our model identifies True Positives. Thus, recall indicates the number of star ratings that we accurately recognized as real star ratings

from all participants. Mathematically, recall is the true positive divided by the summation of the true positives and false negatives.

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

The JRip ML model is explored in detail in this section. According to William et al. [31], JRip is an improvement on IREP in terms of representing 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 approach conventional (i.e., non-incremental) reduced error pruning. The measure used to guide pruning is responsible for the occasional failure of IREP to converge with an increasing number of samples; hence, we improved the IREP metric with a formula, yielding better intuition and fitting behavior.

Table 12 shows that in obtaining the highest numbers of correctly classified instances, Symmetrical Uncertain Attribute Eval, Ranker 10 Attributes with Naïve Bayes Simple, and SMO at approximately 76.92% that utilizes 15 attributes were used. From this, we can see that the highest number of correctly classified instances after attribute selection is achieved by use of Relief F Attribute Eval, Ranker, which uses 10-fold cross-validation and nine selected attributes with 61.54% accuracy for Meta Decorate, Multiclass classifier, and Naïve Bayes Simple. The following charts compare different ML models for the 16 attributes selected.

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

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

Precision is the ratio of true positives to all positives. In our problem statement, the proportion of star ratings was accurately identified as the subjective star rating MOS for all participants. 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).

The recall was used as the next measurement metric, as shown in Figure 7d. This is a measure of how well the model identifies true positives. Thus, recall signifies the number of star ratings 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 3 and 5–8. An analysis of the questionnaires and partial experimental results showed the following:

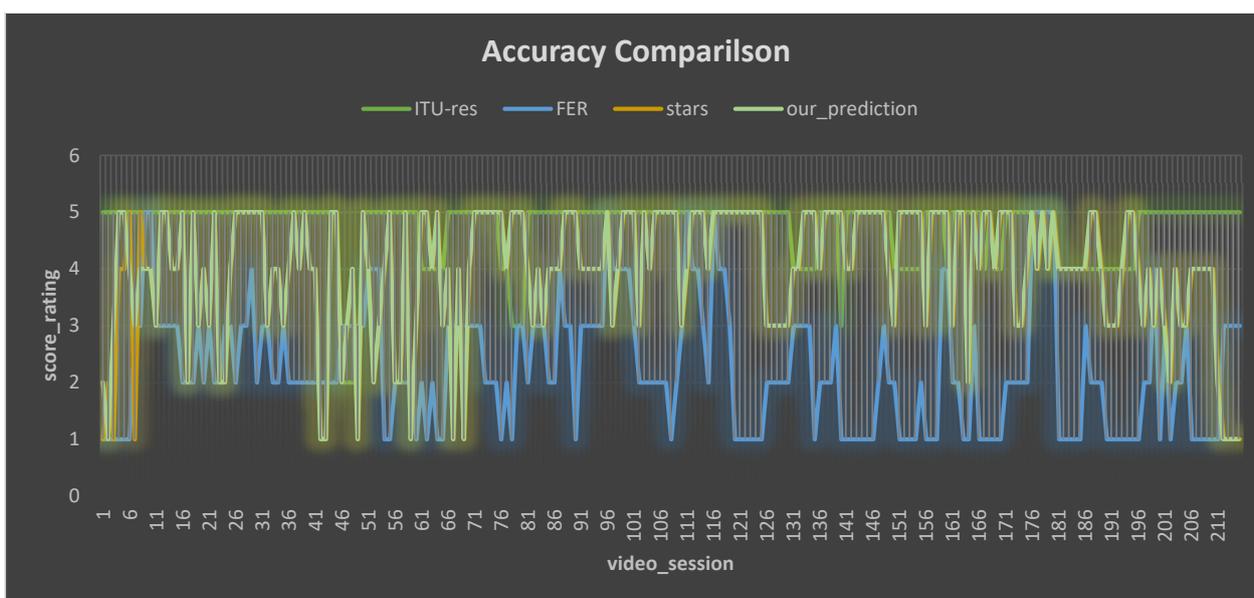
- Massive and intense ads may impact QoE and increase ITU results (i.e., higher bitrate, frame rate, and resolution), but this does not signify that star reviews given by participants will be high;
- Possible QoE IFs from our experimental results include general video content and ad factors (ad length, number of ad, ad location, ad relation to content, repeated ad, and maximum ad acceptance number);
- QoE was most impaired by mid-roll and unskippable ads (approximately 36.2% and 31.9%, respectively). The users found it acceptable to watch an ad of less than 10 s, which is approximately 41.67%.

We extracted the features to be fed into several ML models to obtain better MOS recommendations that would extend the ITU P.1203 standard. The ML models that we

employed to predict MOS by considering features such as FER, ITU P.1203, and advertisement results were tree SPAARC, Random Forest, tree optimized forest, local KNN, multi-layer 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 a class overfitting bias occurred when using 100% training data; therefore, we ruled out using 100% of the training data. Rules JRip and Furia yielded the highest values for the correctly classified instances. Moreover, rule JRip exhibited the highest precision recall.

To answer the next research question on how to provide better recommendations on MOS in order to extend the ITU P.1203 standard, we extracted the features to be fed into several ML models. The ITU P.1203 result can only predict 86 instances correctly from 216 instances, or 39.8% correctly classified instances. From this experiment, including 43 participants, 216 instances, and 16 attributes, we found that it is possible to improve the number of correctly classified classes by 37.12% compared to the ITU-T P.1203 standard results by considering the FER, video metadata, and advertisement data, as shown in Table 11. For all the testing types, we considered only the best values for all measured variables.

Comparisons between the results for ITU-T P.1203, face emotion recognition, and our proposed method are shown in Appendix C, Table A2, and Figure 8.



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

## 5. Discussions and Future Directions

In this section, we discuss related matters and some important details that we believe will be important in the future:

1. Real live FER system—We tried real-time live system face emotion recognition in the real world, and found that it works effectively, although the emotion results do not yet drive the shape of traffic. The proposed method uses emotion-aware advertisement insertions. The shaping of traffic based on emotions to improve QoE will be the focus of future research;
2. Computational complexity of the system—The computational complexity of the proposed method is dominated by facial emotion recognition (FER). The FER process involves several steps including face detection, which involves identifying the location of the face in a video frame. The computational complexity of face detection depends on the deep-face algorithm. The complexity can be written as  $O(n)$ , where  $n$  is the

- number of pixels in a video frame. Feature extraction—This step involved the extraction of features from the face. The computational complexity of the feature extraction depends on the specific features used. Therefore, the complexity of a compound is  $O(f)$ , where  $f$  denotes the number of extracted features. 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 classifier that is used. However, it can generally be considered as  $O(c)$ , where  $c$  is the sum of the emotion classes;
3. Therefore, the overall computational complexity of the FER process can be considered as  $O(n \cdot f \cdot c)$ . Using the proposed method, the FER process was performed on each video frame. Therefore, the overall computational complexity of the proposed method is  $O(T \cdot n \cdot f \cdot c)$ , where  $T$  denotes the total number of video frames. For a video that is 30 s long and has a frame rate of 30 fps,  $T = 30 \times 30 = 900$ . If the video frame is  $640 \times 480$  pixels, then  $n = 640 \times 480 = 307200$ . For  $f = 100$ , features were extracted from each face,  $c = 7$ , emotions were classified, and the overall computational complexity of the proposed method was  $O(900 \cdot 307200 \cdot 100 \cdot 7) = 1.6 \times 10^{12}$ . All this results in relatively high computational complexity. However, it is important to note that the FER process can be parallelized using a GPU to reduce the computational cost of the proposed method significantly. It is important to note that the proposed method selects only the most relevant ads for a particular user. Once the most relevant ad is selected, we incorporate its location, type, and time. These ads can then be provided to users without further facial emotion recognition. Therefore, the overall computational impact of the proposed method was relatively small;
  4. Theoretical analysis of the proposed approach—The proposed Machine Learning (ML) approach for video Quality of Experience (QoE) inference, incorporating face emotion recognition, user feedback on ad insertion, and network conditions, was evaluated using a dataset of 50 recorded video streaming sessions. This dataset included viewers' facial videos, network traffic logs, user feedback on ad insertion, and subjective QoE scores. The accuracy of the model was compared for two baselines—one utilizing only network conditions for QoE inference, ITU-T P.1203, and another employing only user feedback on ad insertion. The proposed approach consistently achieved a lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) than the baseline models, indicating superior accuracy in inferring QoE. This can be seen in Figure 8 and Table 9;
  5. Qualitative analysis revealed the model's sensitivity to viewers' facial expressions, particularly joy, surprise, and frustration, which are known indicators of positive and negative QoE. It also learns to identify advertisement placements perceived as disruptive by users by adjusting its QoE predictions accordingly. Moreover, the model effectively utilizes network bandwidth as a critical indicator of potential rebuffering and stalling, negatively impacting QoE. These experimental results convincingly demonstrate the effectiveness of the proposed ML approach in accurately inferring the video QoE. Its ability to integrate facial emotion recognition, user feedback on ad insertion, and network conditions provides a comprehensive understanding of the QoE. All this offers promising scope to improve user satisfaction and network performance in video streaming systems;
  6. Our hypothesis, which maps the extracted emotion to ACR and MOS, is based on the studies by Porcu et al. [36] and Martinez-Caro and Cano [49]. Porcu et al. [36] analyzed facial expressions and gaze direction. They achieved 93.9% accuracy by leveraging the k-NN classifier, investigating the possibility of estimating the perceived QoE using facial emotions and gaze movement. Martinez-Caro and Cano [46] utilized the ITU-T P.1203 model to estimate MOS values. They used variational algorithms to predict QoE and provided insights into the emotional impact of video quality;
  7. The reasons for choosing ITU-T P.1203 over ITU-T P.1204 in our research were carefully considered in light of the following issues. First, the ITU-T P.1203's requirements are

within our defined parameters, including a 1080p resolution and an H.264 codec. We experimented with these lower resolution and codec standards to simplify video data collection storage and analysis. Second, ITU-T P.1204 is more versatile and generally applicable to multimedia, but we required an assessment standard that is only sufficient for videos. Finally, the main objective of this study was to analyze the impacts of advertisements and video quality on user emotions. Therefore, we simplified this process by using the simpler quality assessment standard ITU-T P.1203. This is because ITU-T P.1204 requires more complex operations and pre-processing, which is beyond the scope of this study. Our study, which investigates the link between video content and viewer emotion, necessitates a specific video quality assessment standard that aligns with our research parameters. We deliberately selected ITU-T P.1203 because of its several advantages within the context of our study design, including direct compatibility issues, established expertise, focus on video quality, the technical expertise of ITU-T P.1203, and the industry standard for video quality. First, direct compatibility (ITU-T P.1203) was specifically designed for 1080p H.264 video content, eliminating the need for complex transcoding or compatibility adjustments. This ensures a smooth and efficient video analysis within the defined parameters. Second, expertise was established. Utilizing P.1203 enabled us to leverage the readily available resources and establish research practices. This enabled our team to focus on core research questions rather than expending resources to adapt and master different standards. Third, our focus was on the video quality. Although ITU-T P.1204 offers advanced capabilities, including support for higher resolutions and codecs, these features are beyond the scope of this study. Expanding the analysis to encompass diverse resolutions and codecs would introduce unnecessary complexity and potentially dilute the focus on our research objectives, which aim to understand the emotional impacts of video content within a well-defined parameter set. Fourth, ITU-T P.1203 has technical expertise in this field. ITU-T P.1203 incorporates sophisticated models and algorithms that simulate human visual perception by considering factors such as spatial and temporal characteristics. This focus on perceptual quality aligns perfectly with our research goals, ensuring that the obtained video quality measurements are directly related to viewers' experience. The fifth reason relates to the industry standard for video quality. ITU-T P.1203 has seen significant adoption as a standardized method for video quality assessment in the industry and research communities. This ensured the consistency and comparability of our findings across different studies and implementations. Therefore, ITU-T P.1203 is the optimal choice for our video emotion research project. Its direct compatibility, established expertise, focus on video quality, and industry-standard implementation ensure efficient, reliable, and relevant analysis within the scope of this study. This selection allowed us to delve deeply into the emotional impact of video content, contributing significantly to the understanding of viewer experiences in current video streaming practices. While we acknowledge the potential value of ITU-T P.1204 to future research endeavors encompassing broader video content types or requiring analyses beyond the current limitations, P.1203 offers a balance of efficiency, expertise utilization, and focused analysis that is optimal for this specific research project;

8. We acknowledge the presence of social elements in our study, particularly by observing user behavior and emotions. However, we firmly believe that this work presents significant research and technical challenges and contributes to video streaming QoE prediction. First, our approach utilizes automated facial emotion recognition (FER) algorithms, moving beyond subjective reports and providing an objective measure of user experience during video streams. This technical approach aligns with research exploring the link between facial expression and emotional responses to QoE. Second, is machine learning model development. We here went beyond measuring emotions. We trained a machine learning model that leverages these extracted features and other technical data points to predict QoE with enhanced accuracy. This technical

innovation offers a data-driven and generalizable solution for improving the user experience in video streaming. Third, compared with the technical benchmark, we demonstrated the technical efficacy of our approach by achieving a 37.1% improvement in accuracy compared to the established ITU-T P.1203 standard, representing a significant technical advancement in QoE prediction. Moreover, we built our website as a unified platform to investigate video QoE inferences using multimodal input. In conclusion, while the study incorporates social elements, such as user observation, its core contribution lies in developing and evaluating a novel, technically grounded machine-learning model for objective QoE prediction using facial recognition. We believe that this work opens promising avenues for improving user engagement and experience in video streaming services;

9. The practical implications of our findings are profound, particularly in the realm of video streaming services and advertisement placement strategies. By leveraging machine learning, facial emotion recognition, user feedback and ITU-T P.1203 results, our proposed framework offers valuable and tangible benefits for users, advertisers, and network providers. Our proposed framework not only improves accuracy but also sheds new light on the detrimental effects of advertisement on user experience. These new perspectives can inform network administrators and content providers regarding the significance of strategic ad placement to optimize overall QoE. Our study sets a foundation for advancements in user-driven video streaming services and advertisement strategies. By demonstrating the effectiveness of our proposed framework, this study opens the door for not only several real-world applications, but also some future research and development. We have shown that an effective QoE inference framework may enhance user experience, offer face-emotion-driven adaptive bitrate streaming capabilities, raise the possibility of targeted ad insertion based on user emotions, offer better content and ad recommendation systems, and allow improved QoE monitoring for network administrators in response to user emotion, as well as content creation with ad marketing. As future research and development directions, there are more possibilities to be addressed related to advanced emotion recognition models that focus on developing more sophisticated FER models that can recognize a wider range of more subtle expressions. In the future, privacy-preserving techniques to anonymize user data while maintaining effectiveness can be more effectively addressed. We also foresee a better chance to integrate multimodal physiological data that combine FER with other sources, such as heart rate, eye tracking or audio, to provide more comprehensive user experiences;
10. The application and potential adaptation of our proposed framework can be explored as follows. First, in the human–computer interaction (HCI) field, our proposed solution can be integrated into HCI systems to provide more responsive and intuitive interactions, such as in virtual reality employing facial expression-driven systems. Second, emotion-aware learning environments can be developed that interpret students' emotional states and attention during learning to adjust content, pace, and difficulty level. Third, we can foresee more emphatic and personalized care delivery in telemedicine and healthcare. Fourth, we could enhance retail and customer experience through videos that show facial expressions, in order to get better information on customer preferences, satisfaction, and engagement levels. Fifth, automotive companies can integrate facial emotion technology into their vehicles so as to improve driver safety and detect driver fatigue, sleepiness, and distraction. Sixth, artists and content creators can develop interactive immersive art performances with real-time facial expression feedback from users. Seventh, law-enforcement agencies could leverage models that detect deception in surveillance footage.

## 6. Conclusions

This study investigated the effects of video advertisements on the user Quality of Experience (QoE) within the domain of end-to-end encrypted video streaming. Our findings

prove that advertisements significantly degrade QoE, in a manner extending beyond traditional metrics, such as stalling and rebuffering. Users' facial expressions, particularly anger, have emerged as significant indicators of QoE degradation. Furthermore, the established QoE estimation approach, based on standards such as ITU-T P.1203, exhibits limitations owing to its inability to capture user emotions and real-time video quality fluctuations.

We addressed these limitations by offering a novel machine learning (ML) framework that utilizes a synergistic combination of facial emotion recognition (FER), advertisement data, and ITU-T P.1203 results in real time. This data-driven approach significantly outperformed the existing methods in terms of accuracy, achieving 37.1% accuracy before attribute selection and 21.74% accuracy after attribute selection, compared to the ITU-T P.1203 standard. By conducting this research, we propose an answer to the research question of how to provide network administrators with insights or predictions about QoE in the context of end-to-end encrypted content in their network, with an accuracy of 76.92% and recall of 0.962%, by considering facial emotions, advertisement data, and video statistics before attribute selection for 15 attributes; this accuracy is 61.54% after attribute selection for 10 attributes when using meta-decorate, multi-layer perceptron, and naïve Bayes. Based on the experimental results and procedures, we conclude that users were more annoyed when an advertisement was placed in the middle of the content. The maximum tolerable advertisement length indicated by the participants was less than 10 s. Moreover, the most annoying advertisement was an unskippable ad, which was in the middle of the content. Furthermore, according to the ITU P.1203 results, only 86 of the 216 occurrences were properly forecasted, and 39.8% were correctly categorized. Rule Chirp, on the other hand, accurately predicted 121 of the 216 cases (76.92%). In this experiment with 43 participants, 216 occurrences, and 15 characteristics, the accuracy of the ITU-T P.1203 standard was improved by considering the FER, ITU-T P.1203 results, and advertisement data.

The proposed framework provides network administrators with valuable insights into user experiences with encrypted content. This information can be used to optimize advertisement placement strategies, resulting in a good balance between user experience and budgeting.

Our experimental research results enhance QoE assessment by establishing the efficacy of utilizing user emotions and real-time video quality data to improve accuracy. We collected a comprehensive and unbiased dataset appropriate for training and evaluating this QoE estimation approach. Further research should investigate the generalizability of this approach across more diverse user demographics and content types.

**Author Contributions:** Conceptualization, T.S. and M.M.M.; methodology, T.S. and M.M.M.; software, T.S.; validation, M.M.M., S.H. and A.B.; formal analysis, T.S. and M.M.M.; investigation, T.S. and S.H.; resources, A.B., M.M.M. and S.H.; data curation, T.S.; writing—original draft preparation, T.S.; writing—review and editing, M.M.M. and S.H.; visualization, T.S. and A.B.; supervision, M.M.M.; project administration, M.M.M.; funding acquisition, M.M.M. and S.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is funded by the UAEU-ADU Joint Research Grant number 12T041.

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

**Acknowledgments:** Thanks to Paperpal Preflight for the checklist before submission, Grammarly for support in grammar and English writing, and ChatGPT and Gemini for help with brainstorming during our research work.

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

## Appendix A

Table A1 lists all the abbreviations used in this study.

**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 in 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
QoS	Quality of Service
UE	User Experience
Weka	Waikato Environment for Knowledge Analysis

## Appendix B

In this section, we list the questionnaire questions. We plan to elaborate upon the questionnaire results 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 the first video (Title: How This Guy Found a Stolen Car!)? [(A) Relaxed-(B) Stimulated]
5. How bodily relaxed/aroused are you after watching the first video (Title: How This Guy Found a Stolen Car!)? [(A) Calm-(B) Excited]
6. How bodily relaxed/aroused are you after watching the first video (Title: How This Guy Found a Stolen Car!)? [(A) Sluggish-(B) Frenzied]
7. How bodily relaxed/aroused are you after watching the first video (Title: How This Guy Found a Stolen Car!)? [(A) Dull-(B) Jittery]
8. How bodily relaxed/aroused are you after watching the first video (Title: How This Guy Found a Stolen Car!)? [(A) Sleepy-(B) Wide Awake]
9. How bodily relaxed/aroused are you after watching the first video (Title: How This Guy Found a Stolen Car!)? [(A) Unaroused-(B) Aroused]
10. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Relaxed-(B) Stimulated]
11. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Calm-(B) Excited]
12. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Sluggish-(B) Frenzied]
13. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Dull-(B) Jittery]
14. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Sleepy-(B) Wide Awake]
15. How bodily relaxed/aroused are you after watching the second video (Title: First Underwater Farm)? [(A) Unaroused-(B) Aroused]
16. How bodily relaxed/aroused are you after watching the third video (Title: Most Beautiful Building In The World)? [(A) Relaxed-(B) Stimulated]
17. How bodily relaxed/aroused are you after watching the third video (Title: Most Beautiful Building In The World)? [(A) Calm-(B) Excited]
18. How bodily relaxed/aroused are you after watching the third video (Title: Most Beautiful Building In The World)? [(A) Sluggish-(B) Frenzied]
19. How bodily relaxed/aroused are you after watching the third video (Title: Most Beautiful Building In The World)? [(A) Dull-(B) Jittery]

Table A1. Cont.

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

Table A1. Cont.

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

**Table A1.** *Cont.*

- 
81. How pleasant/unpleasant do you feel after watching the third video (Title: Most Beautiful Building In The World)? [(A) Bored–(B) Relaxed]
  82. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Unhappy–(B) Happy]
  83. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Annoyed–(B) Pleased]
  84. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Unsatisfied–(B) Satisfied]
  85. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Melancholic–(B) Contented]
  86. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Despairing–(B) Hopeful]
  87. How pleasant/unpleasant do you feel after watching the fourth video (Title: This is Made of...PEE!)? [(A) Bored–(B) Relaxed]
  88. How pleasant/unpleasant do you feel after watching the fifth video (Title: The Most Unexplored Place In The World)? [(A) Unhappy–(B) Happy]
  89. How pleasant/unpleasant do you feel after watching the fifth video (Title: The Most Unexplored Place In The World)? [(A) Annoyed–(B) Pleased]
  90. How pleasant/unpleasant do you feel after watching the fifth video (Title: The Most Unexplored Place In The World)? [(A) Unsatisfied–(B) Satisfied]
  91. How pleasant/unpleasant do you feel after watching the fifth video (Title: The Most Unexplored Place In The World)? [(A) Melancholic–(B) Contented]
  92. How pleasant/unpleasant do you feel after watching the fifth video (Title: The Most Unexplored Place In The World)? [(A) Despairing–(B) Hopeful]
  93. How pleasant/unpleasant do you feel after watching the 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 the 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-min ads long in mid-roll.
    3. In 5 min of video content, every 1 min, there is one repeated ad.
    4. Same ads repeatedly in the pre-roll, mid-roll, and post-roll.
    5. There are no skippable ads.
- 

### Appendix C

In this section, we list the results that were used to build Figure 8. This table compares the results of ITU-T P.1203, face emotion recognition, our proposed method and, as a ground truth, the star rating given by participants.

**Table A2.** Comparison of results derived from ITU-T P.1203, FER, star ratings and predictions.

ITU-res	FER	Star (Ground Truth)	Our Prediction
5	1	1	2
5	1	2	1
5	1	1	3
5	1	4	5
5	1	4	5
5	1	5	4
5	3	1	3

Table A2. Cont.

ITU-res	FER	Star (Ground Truth)	Our Prediction
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
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

Table A2. Cont.

ITU-res	FER	Star (Ground Truth)	Our Prediction
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	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
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

Table A2. Cont.

ITU-res	FER	Star (Ground Truth)	Our Prediction
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
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

Table A2. Cont.

ITU-res	FER	Star (Ground Truth)	Our Prediction
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
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

- Gutterman, C.; Guo, K.; Arora, S.; Wang, X.; Wu, L.; Katz-Bassett, E.; Zussman, G. Requet: Real-Time Quantitative Detection for Encrypted YouTube Traffic. In Proceedings of the 10th ACM Multimedia System Conference, Amherst, MA, USA, 18–21 June 2019.
- 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*, 2851. [\[CrossRef\]](#)
- Bouraqia, K.; Sabir, E.; Sadik, M.; Ladid, L. Quality of experience for streaming services: Measurements, challenges and insights. *IEEE Access* **2020**, *8*, 13341–13361. [\[CrossRef\]](#)
- Agboma, F.; Liotta, A. QoE-Aware QoS Management. In Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia, Linz, Austria, 24–26 November 2008.
- Streijl, R.C.; Winkler, S.; Hands, D.S. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimed. Syst.* **2016**, *22*, 213–227. [\[CrossRef\]](#)
- 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 J. Sel. Top. Signal Process.* **2016**, *11*, 6–21. [\[CrossRef\]](#)
- 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. In Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, Erfurt, Germany, 31 May–2 June 2017; pp. 1–6.
- Garcia, M.-N.; Dytko, D.; Raake, A. Quality Impact Due to Initial Loading, Stalling, and Video Bitrate in Progressive Download Video Services. In Proceedings of the 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, 18–20 September 2014; pp. 129–134.
- Pereira, R.; Pereira, E.G. Dynamic Adaptive Streaming over HTTP and Progressive Download: Comparative Considerations. In Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE, Victoria, BC, Canada, 13–16 May 2014; pp. 905–909.

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. In Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013), Zurich, Switzerland, 14–18 October 2013; pp. 298–303.
11. Hoßfeld, T.; Seufert, M.; Hirth, M.; Zinner, T.; Tran-Gia, P.; Schatz, R. Quantification of YouTube QoE via crowdsourcing. In Proceedings of the 2011 IEEE International Symposium on Multimedia, Dana Point, CA, USA, 5–7 December 2011; pp. 494–499.
12. Oyman, O.; Singh, S. Quality of experience for HTTP adaptive streaming services. *IEEE Commun. Mag.* **2012**, *50*, 20–27. [[CrossRef](#)]
13. Yao, J.; Kanhere, S.S.; Hossain, I.; Hassan, M. Empirical evaluation of HTTP adaptive streaming under vehicular mobility. In Proceedings of the International Conference on Research in Networking, Madrid, Spain, 14–16 November 2011; pp. 92–105.
14. Ghani, R.F.; Ajrash, A.S. Quality of Experience Metric of Streaming Video: A Survey. *Iraqi J. Sci.* **2018**, *59*, 1531–1537.
15. Porcu, S.; Floris, A.; Atzori, L. Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction. In Proceedings of the 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), IEEE, Paris, France, 19–21 February 2019; pp. 82–87.
16. Akhshabi, S.; Anantkrishnan, L.; Begen, A.C.; Dovrolis, C. What happens when adaptive streaming players compete for bandwidth? In Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video, Toronto, ON, Canada, 7–8 June 2012; pp. 9–14.
17. Zinner, T.; Hossfeld, T.; Minhas, T.N.; Fiedler, M. Controlled vs. uncontrolled degradations of QoE: The provisioning-delivery hysteresis in case of video. In Proceedings of the EuroITV 2010 Workshop: Quality of Experience for Multimedia Content Sharing, Tampere, Finland, 9–11 June 2010.
18. Cohen, W.W. Fast Effective Rule Induction. In *Machine Learning Proceedings 1995*; Elsevier: Amsterdam, The Netherlands, 1995; pp. 115–123.
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**, *33*, 363–374. [[CrossRef](#)]
20. Bermudez, H.F.; Martinez-Caro, J.M.; Sanchez-Iborra, R.; Arciniegas, J.L.; Cano, M.D. Live Video-Streaming Evaluation Using the ITU-T P.1203 QoE Model in LTE Networks. *Comput. Netw.* **2019**, *165*, 106967. [[CrossRef](#)]
21. Callet, P.; Möller, S.; Perkis, A. Qualinet White Paper on Definitions of Quality of Experience. In Proceedings of the European Network on Quality of Experience in Multimedia Systems and Services 2013, Novi Sad, Serbia, 12 March 2013.
22. Amour, L.; Boulabiar, M.I.; Souihi, S.; Mellouk, A. An Improved QoE Estimation Method Based on QoS and Affective Computing. In Proceedings of the 2018 International Symposium on Programming and Systems (ISPS), Algiers, Algeria, 24–26 April 2018.
23. Bhattacharya, A.; Wu, W.; Yang, Z. Quality of Experience Evaluation of Voice Communication: An Affect-Based Approach. *Hum.-Centric Comput. Inf. Sci.* **2012**, *2*, 7.
24. Porcu, S.; Uhrig, S.; Voigt-Antons, J.N.; Möller, S.; Atzori, L. Emotional Impact of Video Quality: Self-Assessment and Facial Expression Recognition. In Proceedings of the 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, 5–7 June 2019.
25. Antons, J.N.; Schleicher, R.; Arndt, S.; Moller, S.; Porbadnigk, A.K.; Curio, G. Analyzing Speech Quality Perception Using Electroencephalography. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 721–731. [[CrossRef](#)]
26. Kroupi, E.; Hanhart, P.; Lee, J.S.; Rerabek, M.; Ebrahimi, T. EEG Correlates During Video Quality Perception. In Proceedings of the 2014 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 1–5 September 2014.
27. Arndt, S.; Antons, J.N.; Schleicher, R.; Möller, S. Using Electroencephalography to Analyze Sleepiness Due to Low-Quality Audiovisual Stimuli. *Signal Process. Image Commun.* **2016**, *42*, 120–129. [[CrossRef](#)]
28. Arndt, S.; Radun, J.; Antons, J.N.; Möller, S. Using Eye-Tracking and Correlates of Brain Activity to Predict Quality Scores. In Proceedings of the 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, 18–20 September 2014.
29. Engelke, U.; Pepion, R.; Le Callet, P.; Zepernick, H.J. Linking Distortion Perception and Visual Saliency in H. 264/AVC Coded Video Containing Packet Loss. *Visual Commun. Image Process.* **2010**, *7744*, 59–68.
30. Rai, Y.; Le Callet, P. Do Gaze Disruptions Indicate the Perceived Quality of Nonuniformly Coded Natural Scenes? *Electron. Imaging* **2017**, *14*, 104–109.
31. 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. [[CrossRef](#)]
32. Bailenson, J.N.; Pontikakis, E.D.; Mauss, I.B.; Gross, J.J.; Jabon, M.E.; Hutcherson, C.A.; Nass, C.; John, O. Real-time Classification of Evoked Emotions Using Facial Feature Tracking and Physiological Responses. *Int. J. Hum.-Comput. Stud.* **2008**, *66*, 303–317.
33. 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. In Proceedings of the 9th ACM Multimedia Systems Conference 2018, Amsterdam, The Netherlands, 12–15 June 2018; pp. 466–471.
34. 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. Available online: <https://www.itu.int/rec/T-REC-P.1203.3/en> (accessed on 28 March 2024).
35. Bentaleb, A.; Taani, B.; Begen, A.C.; Timmerer, C.; Zimmermann, R. A survey on bitrate adaptation schemes for streaming media over HTTP. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 562–585. [[CrossRef](#)]

36. Porcu, S. Estimation of the QoE for Video Streaming Services Based on Facial Expressions and Gaze Direction. 2021. Available online: <https://iris.unica.it/handle/11584/308985> (accessed on 28 March 2024).
37. Roettgers, J. Don't touch that dial: How YouTube is bringing adaptive streaming to mobile, TVs. 2013. Available online: <http://finance.yahoo.com/news/don-t-touch-dial-youtube-224155787.html> (accessed on 28 March 2024).
38. Seufert, M.; Egger, S.; Slanina, M.; Zinner, T.; Hoßfeld, T.; Tran-Gia, P. A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun. Surv. Tutor.* **2014**, *17*, 469–492. [[CrossRef](#)]
39. Barman, N.; Martini, M.G. QoE Modeling for HTTP Adaptive Video Streaming—A Survey and Open Challenges. *IEEE Access* **2019**, *7*, 30831–30859. [[CrossRef](#)]
40. Wang, W.; Lu, Y. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *324*, 012049. [[CrossRef](#)]
41. Seshadrinathan, K.; Soundararajan, R.; Bovik, A.C. Study of Subjective and Objective Quality Assessment of Video. *IEEE Trans. Image Process.* **2010**, *19*, 1427–1441. [[CrossRef](#)]
42. Im, S.-K.; Chan, K.-H. Dynamic estimator selection for double-bit-range estimation in VVC CABAC entropy coding. *IET Image Process.* **2024**, *18*, 722–730. [[CrossRef](#)]
43. Chan, K.-H.; Im, S.-K. Using four hypothesis probability estimators for CABAC in versatile video coding. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–17. [[CrossRef](#)]
44. 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. In Proceedings of the 7th International Conference on Multimedia Systems, Klagenfurt, Austria, 10–13 May 2016; pp. 1–12.
45. Han, L.; Yu, L. A Variance Reduction Framework for Stable Feature Selection. *Stat. Anal. Data Min. ASA Data Sci. J.* **2012**, *5*, 428–445. [[CrossRef](#)]
46. Strobl, C.; Boulesteix, A.L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinform.* **2008**, *9*, 307. [[CrossRef](#)] [[PubMed](#)]
47. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347. [[CrossRef](#)] [[PubMed](#)]
48. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC Bioinform.* **2009**, *10*, 213. [[CrossRef](#)]
49. 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](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.