STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, January 2026, pp 629–645.
Published online in International Academic Press (www.IAPress.org)

Predicting the QoE in Adaptive Video Streaming Using Weighted Averaging Federated Learning

Jaafar Rashid, Abolfazl Diyanat*

Department of Computer Engineering, Iran University of Science and Technology, Iran

Abstract Recently, assessing the QoE in adaptive video streaming systems has become an interesting area of research since it directly gauges customer satisfaction. Quality of Experience (QoE) models potentially suffer from data availability issues and a lack in preserving sensitive data of clients. The ITU-T standards provided a guideline for low-cost practical objective QoE assessment to obtain the streaming, non-streaming parameters, and their corresponding accurate MOS scores. The existing QoE prediction literature came with a flavour of implementing the models separately or sharing the QoE data between distinct devices. This work simulating the user interaction with the online video distributors, and obtaining the labelled QoE data using the ITU-T P.1203. In addition, it proposes the Weighted Averaging Federated Learning (WAFL), an enhanced federated learning implementation using the feed forward neural networks to predict the QoE. The WAFL preserves the recent user-privacy requirements by avoiding sharing the entire data among the distributed models and enhance the prediction performance. The training is implemented in a sequential manner amongst the collaborated nodes and enhances the global model by aggregating the learned weights feedback during iterative learning rounds. The achieved QoE prediction accuracy is compared to the single machine learning and the traditional federated learning. The proposed QoE prediction provided an accuracy of 96.06% in estimating the QoE using a small number of streaming and non-streaming QoE parameters.

Keywords Quality of Experience prediction, Federated Learning, Adaptive Video Streaming, QoE Prediction

DOI: 10.19139/soic-2310-5070-2690

1. Introduction

Video on Demand (VOD) services, IP TV, and other video-streaming services are growing rapidly. These services force new challenges for network operators and service providers due to the increased load on the network resources. On the other hand, the users of these services show more user-specific requirements while receiving and playing these video services [1]. Providers compete with each other to satisfy these raised user expectations. Thus, there is a need to understand the relationship between user-experienced service and the parameters of the video and network links. Early efforts have been presented by ITU-T studies toward developing standardized methods to model the quality of experience QoE in video streaming environments. Such standards and recommendations can guide QoE monitoring, assessment, prediction, and proactive management since they are trained and validated on huge-size subjective tests [2].

According to the ITU-T definitions, QoS measures the performance of the delivery infrastructure, including third-party or internal CDN, usually tracking data like overall throughput, latency, error rates, and cache hit ratio. In contrast, the complete end-to-end QoE is defined as “the overall acceptability of an application or service, as perceived subjectively by the end-user”. Mapping between QoS and QoE metrics is highly complex, as they often lay in high-dimensional spaces that make obtaining a closed-form modeling not practical [13].

*Correspondence to: Abolfazl Diyanat (Email: adiyanat@iust.ac.ir). Department of Computer Engineering, Iran University of Science and Technology, Iran.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1.1. QoE Assessment in Video Streaming Environments

Recently, a major of the overall Internet traffic is video streaming. However, due to the sudden fluctuations in bandwidth and many reasons for network performance degradation, it is still a challenging task to stream these videos while satisfying the user's expectations. According to the VNI by Cisco, the predicted ratio of video traffic can exceed 82% of all IP traffic on the internet. HTTP Adaptive Streaming (HAS-based) is a decentralized nature common protocol that allows encoding videos at different representation levels (spatial/temporal/quality) [3]. The video is then divided into chunks (also referred to as segments) of equal duration on the video server. When a client makes his first request, the server responds with information such as duration, segment size, representation levels, or codec type. Accordingly, the client predicts the current bandwidth and buffer status and requests the next part of the video segment with the appropriate quality level. However, the QoE, can also influenced by many factors such as quality switching, initial buffer level, and segment duration. Generally, the QoE influencing factors can be categorized into [4]:

1. System-related: factors such as bandwidth variation, packet loss, delay, jitter, end-user device configuration, and browser,
2. Context-related: factors such as user location, and streaming purpose (i.e., education, and gaming),
3. User-related: factors related to the user expectations,
4. Content-related: factors related to the video content such as bitrate, resolution, and video quality.

For instance, a work by Mrvelj et al.[5] concluded that factors influencing QoE can be divided into *technical* and *nontechnical*. Technical are factors connected to terminals and networks, and nontechnical are factors that are connected to politics and services provided by cellular operators. However, Parmenter et al.[6] in 2015 mentioned that it is difficult to quantitatively define QoE value through network monitoring since there is no unified criterion for QoE measurement in the domain.

According to the ITU-T Rec. P.10/G.100 Amendment 5, the term 'QoE assessment' can be defined as the process of measuring or estimating the QoE for a set of users of an application or service with a dedicated procedure and considering the influencing factors (possibly controlled, measured, or simply collected and reported). The assessment is categorized into two main categories; subjective and objective. To acquire QoE data, Subjective Assessments are conducted either in a controlled environment/lab or with crowdsourcing where influencing factors (input variables) of QoE are monitored together with the user-perceived quality (ground truth labels) forming a database of subjective quality perceptions. In a ground truth scenario, several subjects are asked to rate the visual quality of different streaming videos. The average of these subjective judgments, i.e. mean opinion score (MOS), is computed for the resulted quality measurement, which is usually known as the "ground truth". In the data collection/readiness stage, raw data comprising of input variables and ground truth are acquired and processed further to remove missing entries (optional) followed by the data partitioning to split data into training and testing subsets. The QoE features of HAS (number of rebuffering events, stalling durations, etc.) are extracted/selected from the input variables for both training and testing phases by the ML that can be used for prediction, monitoring, and measurement solutions in QoE management scenarios.

A recent survey by Kougioumtzidis et al.[7] in 2022 mentioned that the "Subjective Assessment" methodologies rely on receiving information from human assessors, who are subjected to a variety of tests or stimuli, while the "Objective Assessment" models on the contrary, can be seen as the mean for evaluating QoE based exclusively on objective quality metrics. The author also showed that the QoE influencing factors (IFs) have been described as the factual condition or adjustment of every feature of a user, system, service, application, or context, that can affect the user's experience quality. The IFs include, among other things, the application or service's type and characteristics, the usage context, the accomplishment of the user's expectations for an application or service, the cultural background of the user, the socioeconomic aspects and psychological portrait of the user, and finally, the emotional condition of the user. Objective Assessment uses the image and video properties and tries to predict the quality as would be perceived by the human. Depending on the amount of source information required, objective assessment can be further classified into;

- FR: full source information required to estimate the QoE.
- RR: partial source information required to estimate the QoE.

- NR: no source information is required to estimate the QoE.

A popular method to evaluate subjective QoE metrics is to ask the user to provide feedback after viewing a video. The users are asked to watch a set of videos and rate each one of them using the MOS scale. The Mean Opinion Score (MOS) is the most popular subjective metric measurement scale that is often used to quantify these factors. Users watch videos and rate them on a five-point discrete scale: 1 (bad), 2 (poor), 3 (fair), 4 (good), and 5 (excellent). The MOS as a subjective metric has become the de facto standard for subjective assessment. It is, however, not easy to automate the MOS measurement since the influence of human psychological factors and user bias needs to be considered. To predict the MOS, a good understanding of the psychology of users to predict the MOS ratings is necessary. These types of studies can measure the effect of the objective metrics as well as the confounding factors on the QoE. The users who are asked to rate the videos could be volunteers such as friends, colleagues, and acquaintances or paid personnel (crowd-sourcing). Such studies are performed in a controlled environment that enables the researchers to have complete control over the stream settings and the user interface. These types of studies can be easily extended to any type of video streaming service as they depend only on the user's feedback.

1.2. Problem Statement, Scope, and the Contribution of This Work

Machine learning-based QoE prediction models that trained on the existing offline QoE datasets might suffer from limited implementation video environments and impracticality, in addition to the limited participant profiles and votes. This prevents models from becoming generic and efficient in predicting the QoE in real video streaming environments. These offline QoE trained models may perform poorly in the test phase on unseen real user data. One the other hand, most of QoE datasets potentially contain user-sensitive information that can bring another challenge in training the QoE models. Thus, there is a need to proposing light weight and standard QoE data obtaining to enhance training the QoE models. In addition, the user-privacy must be preserved to avoid sharing personal information of viewers and decreasing the cost of data transfer with the video streaming networks. Generating the QoE date efficiently according the well-known standards and sharing only the learned weights instead of complete users' data using Federated Learning motivated this work, in addition to insufficient existing research efforts devoted to implementing the FL in QoE modeling.

The main contributions of this research have been listed below;

1. It combines the Federated Learning with the objective QoE prediction, and proposes a specific weighted-averaging to enhance the QoE prediction performance. In addition, it avoids the single point failure of traditional single machine QoE prediction.
2. It keeps the consistency with subjective QoE assessment by implementing the ITU-T standards in obtaining the QoE labeled data.
3. It uses a low-cost ITU-T P.1203 as a truth to evaluate the proposed models to provide a well-tested model capable to perform well on the next unseen real QoE data in real world streaming environments.
4. It considers user privacy and reduce data transfer of QoE prediction by implementing a proposed federated learning approach that enhances the accuracy and convergence.

2. Related Work

Several selected QoE factors are used by Vasilev et al.[13]. The author proposed a Bayesian Network to predict the QoE degradation by the re-buffering ratio and stalling patterns, and implements a sequential training procedure in which the trained model (i) is shared with the next model (i+1) to continue training the next data, then the model is tested on test data at the end of each complete round.

Porcu et al.[11] noticed that it is not proper to merge data of heterogeneous clusters of users since it can reduce the representativeness of unique users in the used QoE model. Thus, based on chosen user-related QoE parameters and rating scores, the author clustered users from a public QoE dataset and then implemented a Cluster-Based

Federated Learning to predict the QoE. K Gao et al.[12] used an end-device probe application to collect QoE parameters, and proposed clustering users according to three user-related temporal parameters (extracted by DL) for FL-based QoE prediction. The author highlighted the role of user-related features in improving the personalized QoE.

Nguyen et al.[14] proposed a blockchain-based FL implementation to avoid the single failure point at the server, and used LSTM and RNN on the QoE parameters of the Poqemon dataset to cluster and classify the observations. With this blockchain setup, the author allows chain devices to vote for model updates and uses a contribution-based vote power mechanism to ensure the objectivity of the resulting aggregated voting.

Xu et al.[18] proposed joining the FL with additional final personalized learning at the clients to enable clients' local model personalized-dealing with network environment changes. The author initially used FL to train a global model based on the clients' weights, then implemented an adaptation phase to train a personalized model for each client to maximize their QoE customization.

Some of the most important related works in this field are summarized in Table 1 along with their characteristics.

Table 1. Comparing the existing Federated Learning-based QoE Predictions

Ref.	YEAR	ITU-T Standards	DATASET	QoE PARAMETERS	ML	Performance Metrics	FINDINGS	LIMITATIONS
[17]	2025	✗ No	multi-node simulation	user context, network conditions, and device capabilities	ML	Accuracy and Bandwidth consumption	FL on multiple nodes can enhance the accuracy and decrease the consumed bandwidth	no ITU-T support, no video parameters
[19]	2023	✗ No	Broadband and 4G LTE datasets	chunk quality level, Quality difference between consecutive chunks, and Rebuffering time	Deep reinforcement learning (DRL)	Average rewards, Convergence rate, Variations in rewards across different bandwidth environments	combining FL with DRL can enhance QoE models	no ITU-T support and complexity

Continued on next page

Ref.	Year	ITU-T Standards	Dataset	QoE Parameters	ML	Performance Metrics	Findings	Limitations
[10]	2019	✗ No	Web Browsing QoE Subjective Test Dataset V 1.0	max downlink bandwidth, browsing time until the rating prompt (surfing duration), and time consumed during the rating process (prompt duration).	CL, FL, and RRL	Training Time, and AUC ROC.	CL can outperform the IL in prediction performance and data privacy	Limited parameters, focused only on web contents, no video parameters, no ITU-T support, only binary MOS prediction is implemented.
[11]	2022	✗ No	Poqemon-QoE	Clustered users according to user-related parameters (age, sex, and study level).	K-mean clustering, and NN classification.	P-Correl., Accuracy, Recall, and F1-Score.	Clustering the FL data according to users' age enhances the prediction. Buffering time has the most impact on QoE.	ITU-T support, specific for QoE data contains user-related parameters.
[12]	2020	✗ No	Collected from end-devices application	user-related parameters extracted by DL	K-mean clustering	Scatter diagram and correlation between the actual subjective and predicted MOS.	Personalized QoE can be improved by analyzing the user-related parameters.	no ITU-T support, required end-device prob application.
[14]	2023	✗ No	Poqemon-QoE	-	RNN, LSTM	MSE	Blockchain-based FL can solve the single failure point issue.	no ITU-T support, no clear video parameters.
[18]	2023	✗ No	Emulated environment using Google Chrome and Apache server	Bitrate, rebuffering time, and bitrate variation.	RNN, LSTM, and CNN	CDF and Reward functions	Joining FL and client personalized learning can enhance the QoE prediction	no ITU-T support, complexity

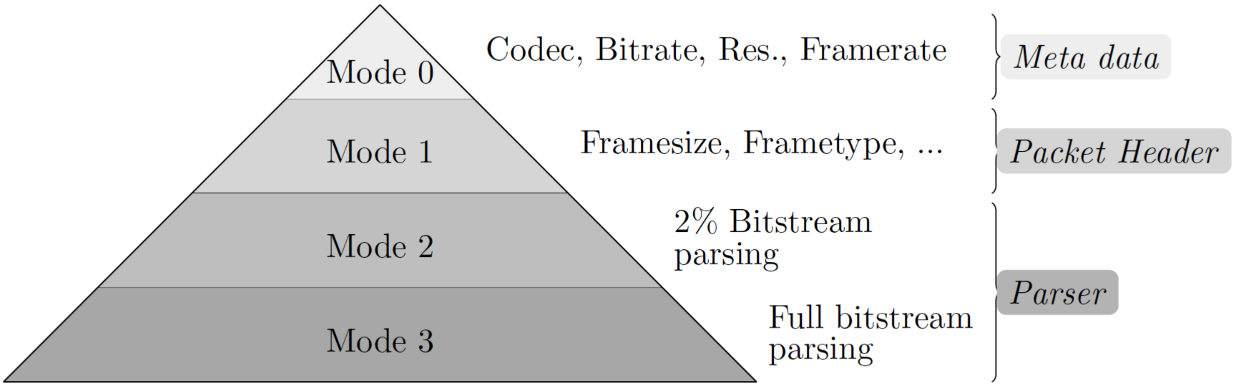

Figure 1. The four operational modes of accessing data in ITU-T P.1203 rec.

3. Background

3.1. *QoE Standard models for adaptive video streaming*

The ITU-T P.1203 (formerly called P.NATS) was published in 2017 as an early standardized audiovisual quality model for HAS services. It predicts the user's QoE for sequences of up to five minutes in length[8] . The primary purpose of this model was to monitor the transmission quality for operators and maintenance specialists since it can be deployed both in end-point locations and at mid-network monitoring nodes and devices. This standard adopted predicting the QoE as a MOS on a 5-point ACR scale. In addition, it allows the reporting of several QoE-related diagnostic indicators. This model allowed a better understanding of the relation between QoE degradations and parameters like quality switching of bitrates, resolutions, and frame rates, in addition to the initial loading delay and stalling. For flexible implementation, the P.1203 standard came with four modes of operation regarding the input stream shown in Figure 1.

In Mode 0, Information is obtained from meta-information available during progressive download or adaptive streaming, such as information from manifest files used in DASH about codec and bitrate, and initial loading delay and stalling. Mode 1, uses all information from Mode 0 with additional video and audio frame information based on packet header inspection. In Mode 2, all information from Mode 1 and up to 2% (in Bytes) of the overall media stream information obtained by deep packet inspection and partial bitstream parsing. The last mode (Mode 3), uses all information from Mode 2, and complete media stream information based on bitstream parsing. For instance, the QoE degradations can relate to parameters such as;

- initial loading delay, and stalling, that are highly related to rebuffering at the client.
- media adaptations that refer to events where the player switches video playback between a known set of media quality levels (such as; bitrate, resolution, frame rate, and similar attributes) while adapting to network conditions.

The ITU-T also tried to clarify the definitions lying under the umbrella of QoE modeling such as;

- play-out: The process of transferring buffered information to a player, such as the frame information of a video stream.
- play-out buffer: Transient memory where the video is stored for play-out.
- stall: A condition occurring at the presentation layer, when media play-out is suspended. This condition typically occurs due to play-out buffer depletion. Stalling is caused by rebuffering events on the client side, which could be a result of video data arriving late. Usually, rebuffering events are indicated to the viewer, e.g., in the form of a spinning wheel, and result in stalling of the media playout.

- re-buffering: A condition occurring in the media buffer when the fill level is sufficiently depleted and buffer exhaust is imminent. This condition typically results in a suspended video play-out.
- initial buffering: A condition occurring in the media buffer on the initiation of a media stream, and completing when the configured buffer fill level is accumulated. This condition typically occurs before the start of media play-out.
- media adaptation: Events where the player switches video playback between a known set of media quality levels while adapting to network conditions, by downloading and decoding individual segments in sequence.

Such standardized definitions and calculations encouraged academic research to analyze the QoE parameters and understand the objective relation between the QoE and its influencing factors. Since the ITU-T standards have been evaluated on huge data and the objective estimated QoE levels have been compared to extensive subjective tests, it became a useful guide to model the QoE in different environments and experimental tests.

3.2. Content Privacy Vs. Content-Oriented Privacy in video streaming

Regarding privacy in video streaming, ‘Content Privacy’ focuses on protecting the actual content (video itself and associated metadata) being consumed by users from unauthorized access and distribution (i.e., Encryption, Access Control, authorization, data leakage prevention). On the other hand, ‘Content-Oriented Privacy’ focuses on protecting the privacy of user data and preferences related to the content they consume (i.e., user interactions, viewing habits, Data Anonymization, viewing history, preferences, and interaction patterns.[9]

3.3. Federated Learning in QoE Prediction

Machine Learning algorithms can be trained collaboratively, where the ML model can be trained on separated data lakes and contribute to the final model. This partial training on isolated data sets then allows combining or averaging the weights and using them to enhance the final model using various collaborative and federated learning techniques[10]. However, collaborative learning might bring privacy issues while sharing sensitive user data. On the other hand, Federated Learning (FL) allows storing the user’s raw locally in the user’s device for partial training, while sharing the obtained weights from the trained models to aggregate them in the central server. The multiple entities collaborate in solving a machine learning problem, under the coordination of a central server by sharing only the updates intended for immediate aggregation[11]. Thus, the FL was the key to sharing the user-related data to resolve the data sparsity (data sparsity often occurs since observable data from the individual user is not always sufficient) issue while protecting the user’s privacy. Thus, FL allows the following;

❶ learning from decentralized data sources without the need to transfer the raw data.

❷ keeping data on local devices and only sharing model updates or gradients.

❸ enhancing privacy, security, and efficiency since it reduces data transfer, and lowers bandwidth and storage requirements. It also reduces risks associated with data centralization and enhances security by avoiding sensitive data sharing.

The general step of FL can be summarized by the following;

❶ A shared global model is sent to local devices.

❷ Iteratively, each device trains the model using its local data.

❸ Devices then send the model updates (gradients) back to the central server.

❹ The server aggregates the updates from multiple devices to improve the global model.

❺ Multiple rounds of local training and aggregation are repeated until convergence.

In the QoE prediction context, the FL can generally implemented by; Collecting video parameters and corresponding MOS scores from distributed devices, Preprocessing the data to ensure consistency in format and quality, and Distributing a global prediction model t the devices. Then, each device uses its local video parameter and MOS scores to train the model locally, each device computes the model updates (i.e., weights and bias) and sends them to the server. The server in his role, aggregates the updates (e.g., by averaging) to update the global model.

4. Methodology and Dataset

4.1. Obtaining the Data using ITU-T Standards

This phase inspired by a previous work[16] by our QoE team. Short HD videos of H.256 encoding are initially requested from the 'Aparat.com' website. Selenium is used to simulate the user actions of searching, selecting, and playing the video content. A portion of each video (specific segment) is viewed, and its labelled ITU-T P.1203 QoE parameters are extracted from the HTTP Archive (HAR) file and the time stamp files. In addition, network parameters have been collected during viewing sessions of each video. The extracted QoE parameters and MOS labels are then stored in a structured database as numerical tabular data. The ITU-T P.1203 Mode 0 is used to keep the simplicity in this work since it has access to the least amount of Meta-data (Encrypted media payload and media frame headers). Video and network parameters are extracted to be used as predictors to train ML algorithms. Predictors used in this work include;

- Delay: the time to transfer data from source to destination, which is an end-to-end time usually measured in milliseconds. A long delay can cause laggy interactions, such as an unresponsive play button, or negatively affect the delivered live streaming service.
- Throughput: represents the amount of data transmitted per unit time (e.g., Mbps) through the network (actual data rate during the transmission). Low throughput less than the required video content bitrate often results in low-quality playback and annoying rebuffering, and interrupts the video smoothing at the client side.
- Connection time: the required period to establish the connection between the client device and the server. Longer connection time affects session start and provides a bad first impression even before playing the content.
- Time to First Byte (ttfb): the time from the client's request initiation to the first byte received. Long ttfb reflects server or network inefficiency and impacts user trust in service quality.
- Startup time: the time from video request until playback starts, which is one of the important QoE influencing factors. For instance, a client might start abandoning video after a delay of more than 2 milliseconds.
- Buffering time: the time spent in the client-side buffering during video playback. Longer buffering time can cause annoying interruptions at the client side.
- Buffering ratio: the ratio of the total playtime users have experienced buffering for during a video session. The ratio is obtained by dividing the buffering time by the total playtime. A high ratio (e.g., 20% of video is buffering) is often associated with a bad subjective QoE score by clients.
- Avg rebuffering time: the average duration of buffering events, where longer rebuffering is more annoying and often impacts the smoothness.
- Avg bitrate: the average number of bits per second of the streamed encoded video. Higher bitrate allows better visual quality and larger size data transfer. Frequent switching in bitrate by ABR algorithms can also annoy the client due to the imbalanced bitrate with network conditions.
- Video width / video height: are dimensions that decide the resolution of the streamed video. Higher resolution can allow better service quality by using a suitable bitrate. Inconsistent values of resolution and bitrate can cause blurriness in the watched video.
- Jitter: measures the variability in packet delay over time through the network. Higher jitter can cause unstable playback and unpredictable buffering events.
- Packet loss: the percentage of video data packets lost during transmission. High packet loss impacts the subjective QoE score since it can cause frame drops and artifacts.

- Stalling: the number or presence of playback interruptions (stalling events). Stalling is a QoE degradation that consists of playback interruptions (often caused by rebuffering events).

4.2. Federated Learning for QoE Prediction

After obtaining approximately 4000 labelled sample of QoE data using the ITU-T objective assessment, The obtained data has been used to train a Feed Forward NN classifier, and predict the QoE in different scenarios of parameters' sets. The MOS class labels of decimal numbers ranging from one to five are mapped into five-level categorical classes ranges from 1 to 5. MATLAB R2022b is used to train the model with implementing Federated Learning in multiple rounds, and iterative learned weights feedback to a global central model. A single hidden layer Feed-Forward NN model for classification is trained to learn parameters. ReLu is used as an activation function in the hidden layer, and SoftMax is used to support multiple class labels in the output layer. Initializing the weights with a value (w) to forward propagate an input to the next layer after adding the bias such as;

$$z_l = w_l a_{l-1} + b_l \tag{1}$$

The output is then regularized using the ReLu activation function and forwarded to the next layer;

$$a_l = \text{ReLU}(z_l) \tag{2}$$

The He Initialization[22] is used as an initialization to bring the mean of the activations to zero;

$$w_l \sim \mathcal{N}(\mu = 0, \sigma^2 = \frac{2}{n_{l-1}}) \tag{3}$$

and the biases are initialized to zero:

$$b_l = 0 \tag{4}$$

Here, n_{l-1} represents the number of incoming connections to the neuron (i.e., the size of the previous layer). The normal distribution has a mean of $\mu = 0$ and a variance of $\sigma^2 = \frac{2}{n_{l-1}}$. The 'He Initialization' allows avoiding the inactive neurons issue by ReLU function (when its output is zero due to negative input). Thus, it ensures that more neurons remain active and contributes during the training process.

To measure the the overall effeciency of the model, accuracy has been calculated by the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions as shown in Eq.6.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Samples}} \tag{5}$$

Federated Learning has been implemented to enhance the QoE prediction while protect user privacy. The initial global FFNN model is shared with clients in a low-cost data sharing. Then each client model iteratively trains using its own local data and sent the weight updates back to the central server. The server finally aggregates the clients' updates to improve the global model. The FL is implemented in multiple-round manner of local training and aggregation until convergence.

The proposed network supposes a single server to distribute the initial model to the clients, and training clients C1, C2, C3, ..., CI separately, each on its own labelled QoE data obtained by simulating the ITU-T P.1203. The federated learning has been implemented in J rounds. For comparing the performance, different number of clients, rounds, and different tunned hymperparameters have been used. The implemented federated learning is shown in Fig2.

Regarding the number of clients in FL implementation, the author in [20] observed that the accuracy might decreases when increasing the number of clients. Where in such implementations, the authors use small amount of local data since they divide their collected data among a large number of FL clients. Thus, the training phase will be more prone to over fitting or insufficient training conditions. Thus, in this proposed WAFL, 5, 10, and 15 as logical and practical range of FL clients are experimentally used to explore the impact of FL clients on the QoE prediction

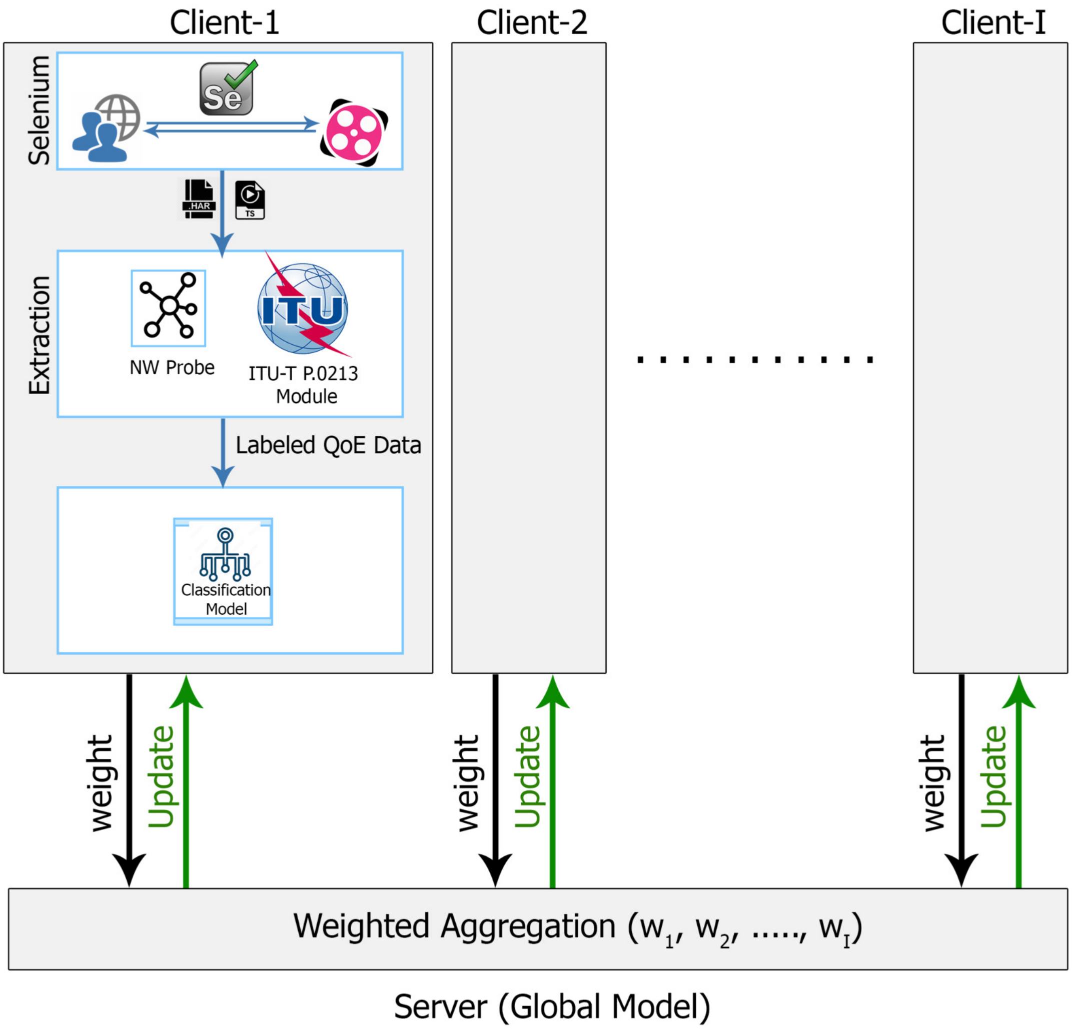

Figure 2. Server-client interaction and Aggregating clients' update of the proposed WAFL QoE prediction

accuracy. The author in [20] also concluded that the accuracy is often increased as the number of classes in each party increases since it enhances the learning pattern in the training phase. In addition, many hyper-parameters can also affect the FL model performance such as; controlling the weight by the L2 regularization lambda that needs to be carefully tuned to achieve good accuracy and convergence. Learning rate can also affect the performance (a typical values 0.1, 0.01, and 0.001 have been used with WAFL to study the impact of learning rate on the proposed model).

Internally, in the iterative ML models such as the used FFNN classifier, the batch size also can impact the performance. Larger batch allows smoother weight update (gradients), more stable convergence, but slower per-round updates. On the other hand, Larger number of local epochs can fasten the local learning but also can amplify divergence under non-IID.

Regarding the number of FL rounds, it is a reasonable number related to the achieved accuracy and the convergence speed. Convergence is often ruled by a specific high accuracy level after which the model training is stopped, or a specific low loss values, or several patience rounds spent without enhancing the accuracy. For example, an accuracy threshold of 98% and several patience rounds of unchanged accuracy might be used as stop conditions.

Regarding to the IID property of the client data used in training the FL model, literature reported that achieving stable learning and fast convergence is still an open problem on non-IID data[20]. However, several solutions are proposed such as initial clustering of clients according their data[11]. However, practically in video streaming and user voting environments, clients often prefer protecting their privacy and avoiding access their local data such what happen in FL implementation. Thus, in this work we consider that the central server does not know about the class distribution (MOS scores of video viewers) and the data has a not-IID property. An additional theoretical analysis has been done to explore the impact of not-IID data on the QoE prediction by preparing a processed IID data extracted from the original real clients' data.

The proposed system also allow handling heterogeneous client computational power since it implement a light weight software to train and test a simple machine learning model on a low size tables of numerical QoE parameters collected by ITU-T module. Thus, it also does not force any heavy communication load since it only transfer the model update instead of sharing the entire client data.

This work also consider that the client selection is done randomly while initiating the communication session between the clients and central server. The number of selected clients are considered to be equal or less than the full participants (video viewers) existing in the video streaming network. Data distribution across clients and distributing the global model to FL clients in this work is simulated using MATLAB R2022b. However, this simulation initially suppose that the client model and its local data lies on the same device or aggregated in a nearby aggregating node on the network between the client and server.

4.3. Weighted Averaging Federated Learning for QoE Prediction

Compared to traditional implementation of federated learning fro predicting the video quality, a weighted federated averaging FL has been proposed in this work to enhance the prediction accuracy and fasten the model convergence. In addition to federated learning advantages such as protecting the user-privacy and enhancing the resource utilization (sharing the light weight update instead of sharing the local data), it also cope the clients' data variety challenge. In real word video service environment, clients have different video requirements and network condition. Thus, data samples used in training the machine learning models is often unfairly distributed among different clients. Traditional averaging of weight updates regardless the amount of samples of each client might add new challenges of biased learning phase, that affects the accuracy and convergence of iterative machine learning models. A specific weight according to the samples amount of each client is proposed as an importance indicator of each client during aggregating the clients' feedback using federated averaging. suppose that the client c_i has a large amount of samples S_i compared to other clients, such client must get more attention during the training and during averaging the updates of clients compared to other poor data clients. Equation 6 shows how can the importance indicator of each client be calculated according to its data size.

$$P_i = \frac{S_i}{\sum_{i=1}^{I} S_i} \tag{6}$$

where, S_i is the number of samples existing in the c_i training data, and P_i is its importance indicator or the proportion of its data amount to the total client's data amount. In the averaging phase at the server side, the clients' update feedback is aggregated using the formula 7.

$$W_g^j = \frac{\sum_{i=1}^{I} P_i W_i}{I} \tag{7}$$

where, W_g^j is the aggregated update sent from clients C_i to the global model in FL round r_j.

In the real world video streaming environments, service providers and network operators often tend to minimize accessing the user's personal ranking and voting information. Thus, this work suppose that the global model on the operator side does not directly see the class distribution of the training data. The collected QoE data using the ITU-T P.1203 module is a Not-IID data since each client has its different video viewing and its network conditions. However, a theoretically balanced data is also prepared to explore the impact of the IID property in training data on the prediction model's performance. A redistributed balanced data is selected from each client's data to train the model and compare the accuracy and convergence with the practical Not-IID data. Different video service clients over the internet often behave differently during playing their favourite videos and voting them due to their different network conditions and their personalized requested services. Thus, the unfairly distributed training data in the amount of samples and in the appeared class labels affect the model learning negatively. The obtained labeld data from ITU-T P.1203 module is divided randomly (Not-IID) into five data files. A global model (a single hidden layer FFNN classifier) has been initiated with He weights, and distributed to five clients in the first FL round. The clients trained the model each on its own data, and sent its feed back update to the global model. The central server aggregate the clients updates using FedAvg to calculate the new weights of the global weights. The update and averaging is continued in multiple FL rounds until meeting the convergence condition. An accuracy threshold of 98% and a patience of 10 rounds are used as stopping condition of convergence. The used class distribution and the number of samples used in exp.2 are shown in 2.

Table 2. The samples' details of FL clients' data used in Exp.2

Client	Class=1	Class=2	Class=3	Class=4	Class=5	#samples
client1	11	274	216	154	8	663
client2	11	307	251	185	10	764
client3	17	379	298	197	10	901
client4	15	336	251	167	9	778
client5	15	347	271	193	10	836

The proposed WAFL for QoE prediction is described in Algorithm 1.

Algorithm 1 Weighted Averaging Federated Learning for QoE Prediction

```
Input:
  • $R$: Number of FL rounds $(r_1, r_2, \ldots, r_J)$
  • $C$: Set of clients $(c_1, c_2, \ldots, c_I)$ selected randomly
  • $W_k$: Initial global model weights
  • $X_i, Y_i$: QoE-labeled data of client $c_i$
Output: $W_g$: Final global model weights
Initialize global model weights: $W_g \leftarrow W_k$
for $j = 1$ to $J$ do                                              ▷ All FL rounds
    for $i = 1$ to $I$ do                                          ▷ All FL clients
        Calculate $S_i$: number of samples of client $c_i$
        Calculate importance of client $c_i$: $P_i = \frac{S_i}{\sum_{i=1}^{I} S_i}$
        Train the client's model locally on $(X_i, Y_i)$
        Calculate the prediction accuracy $Acc_i$
        Obtain $W_i$: updated weights from client $c_i$
    end for
    Aggregate updates using Weighted FedAvg: $W_g^j = \frac{\sum_{i=1}^{I} P_i W_i}{I}$
    Update global model: $W_g \leftarrow W_g^j$
    Stop rounds if Convergence condition is met (Threshold of 98%, Pathience of 10 rounds)
end for
```

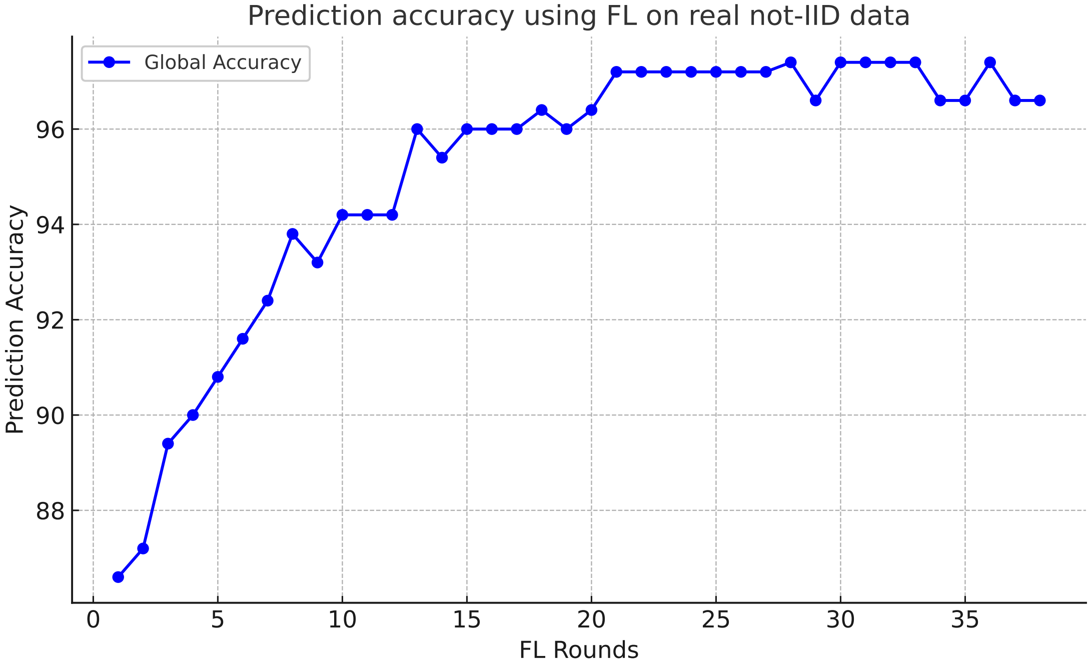

Figure 3. Prediction accuracy using FL on real not-IID data in exp.2

5. Results and Discussion

Initially, the obtained labelled QoE data from the ITU-T P.1203 simulation has been used to train a feed forward neueal network (FFNN) model as a classifier without implanting the federated learning. The classifieier is trained using 80% of the collected data, and tested using the remaining 20% of data. A 5-fold cross-validation is also used during training to tune the hyper parameters. The performance of the QoE prediction has been measured using prediction accuracy as performance metric. A single device model achieved a high accuracy using a single hidden layer, learning rate of 0.001, L2 regularization lambda of 0.001, max epochs of 100, and min batch size of 32. The model's hyper-parameters have been tuned using a 5-fold cross validation, the obtained accuracy during training was 87.55%, which was slightly differ from the test accuracy (test accuracy is 89.95%). The high training accuracy reflects that the model learned well on the used training set, while the high-test accuracy showed that the model can generalize well in handling new unseen data and has a low level of overfitting. Macro-averaged metrics that are used to evaluate systems performance across different datasets (a held-out test data in this work) is also measured for the proposed model. The achieved average macro-Recall of all classes (85.83%) showed that the class of most true test samples have been detected correctly. The achieved F1-score of classes of class 2 = 95.94%, class 3 = 83.64%, class 4 = 75.08%, and class 5 = 90.10%, showed a balanced performance between precision and recall for these classes. The relatively low 75.08% F1-score of class 4 can be considered acceptable due to the low proportion of class 4 (class 4 samples represents 12.87% out of other classes' samples) shown in the randomly selected not-IID test set used in this experiment. In the second experiment, a traditional federated learning (discussed in Section 4.2) has been implemented to explore the impact of FL in QoE prediction. Since the FL implementations generally focus on protecting the user-privacy, a real clients' data obtained from ITU-T module has been used to keep the consistency with the real world environment. The used data was similar to real data regarding the Not-IID property since different clients behaves differently and experiences different network conditions. The achieved results in exp.2 are shown in Figure 3. The results showed that implementing a five-client traditional FedAvg federated learning on Not-IID QoE data can enhance the prediction accuracy as compared to

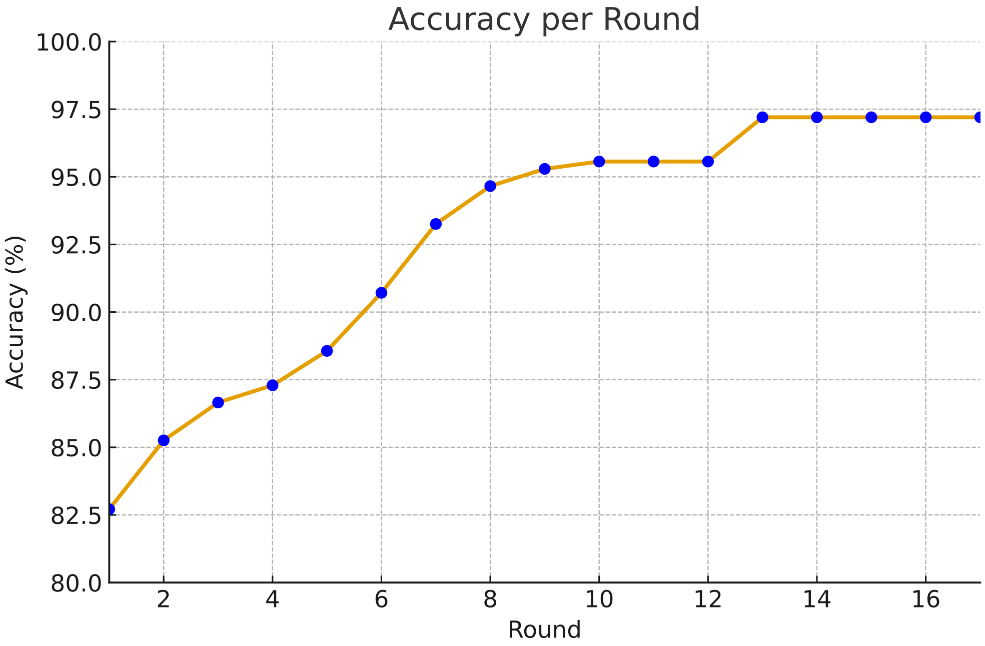

Figure 4. Prediction accuracy using WAFL on real not-IID data in exp.3

implementing a single device model. The achieved accuracy increased form 87.48% (acheived by a single device in exp.1) to an accuracy of 96.60% by implementing federated learning. The model is converged after 38 rounds to achieve its max prediction accuracy. The following parameters are used in implementing the FL; maxRounds = 50, client Local Epochs = 50; mini Batch Size = 20, learning Rate = 1e-3, patience = 10, min Delta = 1e-3, target Accuracy = 98%, reproducibility = rng(0), and hidden layer Size = 64.
In the third experiment, the proposed WAFL is implemented to explore the impact of the proposed weight used in aggregating the clients' update. The results of exp.3 showed an enhancement in the QoE prediction performance after cosidering the clients' data amount in aggregating the updates as shown in Figure 4.

Considering the client data size in aggregating the updates enhanced the QoE prediction accuracy to 97.02% in through only 16 rounds of WAFL (in exp.2 the traditional FL achieved max accuracy of 96.60% after 38 rounds). For further analysis, to explore the impact of learning rate on the prediction performance. The values 0.1, 0.01, and 0.001 have been used to experimentally check its impact on a 5-clients WAFL model on not-IID data. As shown in Figure 6, the results showed that using 0.001 learning rate achieved the highest accuracy and the best convergence time within only 17 WAFL rounds. While using a larger value of learning rate such as 0.1 increased the number of rounds required to meet the convergence conditions to more than 27 rounds with less accuracy of 91.85%.

To address the impact of FL client number, an additional test has been implemented using 5, 10, and 15 FL clients. The results showed that increasing it slightly increases the convergence rounds. However, practically, increasing the number of clients often leads to decreasing the number of samples of each client's data due to dividing the limited data among more clients. Thus, the resulting small size per client data will also decrease the learning and test time. Thus, it can be considered an acceptable impact on convergence as shown in Figure.

The results of implementing the proposed WAFL increased the QoE prediction accuracy with approximately 10% as compared to simply predicting the QoE using a single model implementation. In addition, the advantage of FL in protecting the privacy and avoiding single failure point are also gained.

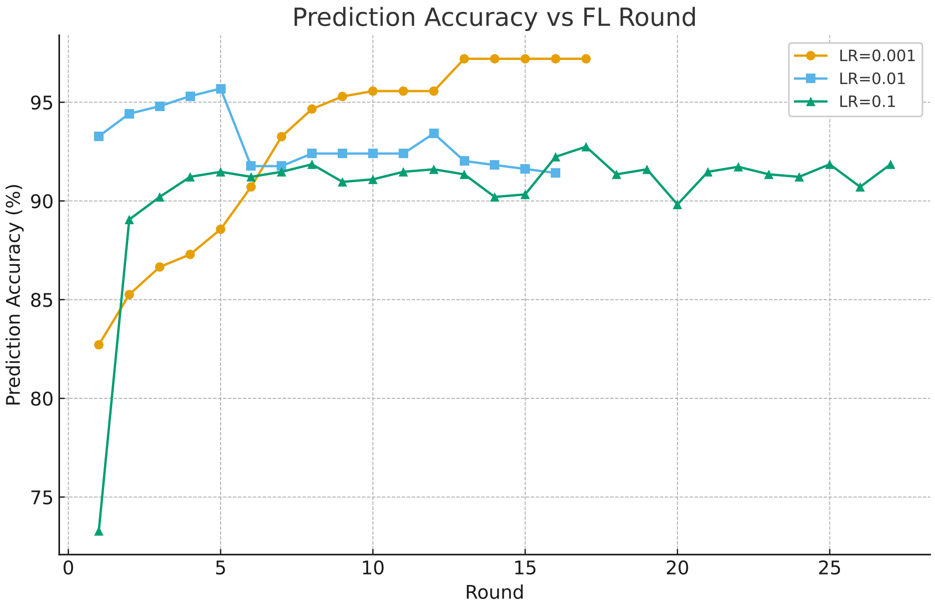

Figure 5. The impact of learning rate on WAFL model

6. Limitations

This work is experimentally evaluated using limited choices of FL clients number of 5, 10, and 15 due to the time limitations of this study. However, more tests consider larger number of clients can also be implemented in the future to strength the study results and prove the scalability of the proposed WAFL model. The Failure in connecting a client during the WAFL rounds and to avoid it using specific reconnect timers or eliminating the disconnected client data impact by redundancy or initial spare clients, is also an aspect that does not be discussed and addressed in this work. The security aspects such as the model robustness against faulty or malicious data, and attacks in federated learning is also out of the scope of this prediction work.

7. Conclusion and Future Work

This article presented a collaborative machine learning approach to enhance the QoE prediction in video streaming. The achieved accuracy using the feed forward NN has been enhanced by adding a weight-based averaging during implementing the federated learning. The results showed how can the WAFL enhance the QoE prediction accuracy while protecting the user data privacy by avoiding sharing their sensitive data. The used supervised learning has been implemented using MOS scores obtained by a light weight ITU-T simulation to enhance the applicability and avoid the high cost of the offline subjective sessions. Although the evaluation of the proposed methods has been performed with only one dataset, it covered a large number of samples and included a randomly selected types of actual online publicly available short video scenes. Using larger size of QoE data generated by ITU-T simulation is scheduled for future work in addition to novel scenarios of combining the wireless telecommunication signal quality parameters of LTE and 5G with the QoE parameters used in this work to explore the impact of these parameters on the predicted QoE score.

This light weight prediction model can be used by network operators and video service providers to monitor the clients sanctification by the well known ITU-T rec., and to proactively obtain a warning report about the next

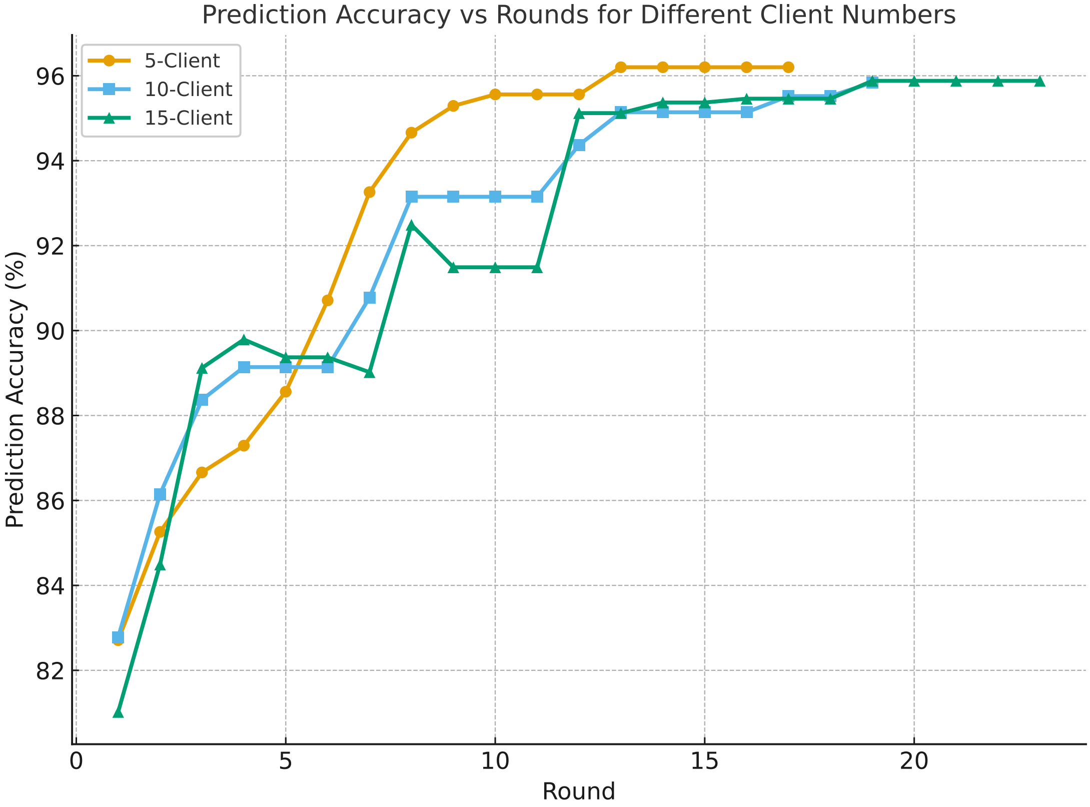

Figure 6. The impact of client number on WAFL model

possible QoE degradation. This aslo allows operators to avoid experiencing bad quality conditions by end users by invoking their prepared load balancing and congestion solving techniques.

Declaration

No funding was received for conducting this study. The authors have no competing interests to declare that are relevant to the content of this article.

REFERENCES

1. M. S. Allen, B. Y. Zhao, and R. Wolski, *Deploying Video-on-Demand Services on Cable Networks*, In 27th International Conference on Distributed Computing Systems (ICDCS '07), pp. 63–63, 2007.
2. K. Bouraqia, E. Sabir, M. Sadik, and L. Ladid, *Quality of Experience for Streaming Services: Measurements, Challenges and Insights*, IEEE Access, vol. 8, pp. 13341–13361, 2020.
3. A. Biernacki and K. Tutschku, *Performance of HTTP video streaming under different network conditions*, Multimedia Tools and Applications, vol. 72, no. 2, pp. 1143–1166, 2013.
4. D. Vučić, S. Baraković, and L. Skorin-Kapov, *Survey on user perceived system factors influencing the QoE of audiovisual calls on smartphones*, Multimedia Tools and Applications, vol. 82, no. 16, pp. 24681–24706, 2022.
5. Š. Mrvelj, M. Matulin, and S. Martirosov, *Subjective evaluation of user quality of experience for omnidirectional video streaming*, Promet–Traffic & Transportation, vol. 32, no. 3, pp. 409–421, 2020.
6. D. Parmenter, *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs*, John Wiley & Sons, 2015.
7. G. Kougioumtzidis, V. Poulkov, Z. D. Zaharis, and P. I. Lazaridis, *A survey on multimedia services QoE assessment and machine learning-based prediction*, IEEE Access, vol. 10, pp. 19507–19538, 2022.

8. W. Robitza, S. Göhring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, and S. Broom, *HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203*, In Proceedings of the 9th ACM Multimedia Systems Conference, 2018.
9. Q. Wu, Z. Li, G. Tyson, S. Uhlig, M. A. Kaafar, and G. Xie, *Privacy-aware multipath video caching for content-centric networks*, IEEE Journal on Selected Areas in Communications, vol. 34, no. 8, pp. 2219–2230, 2016.
10. S. Ickin, K. Vandikas, and M. Fiedler, *Privacy preserving QoE modeling using collaborative learning*, In Proceedings of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks, pp. 13–18, 2019.
11. S. Porcu, A. Floris, and L. Atzori, *CB-FL: Cluster-Based Federated Learning applied to Quality of Experience modelling*, In 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 585–591, 2022.
12. Y. Gao, X. Wei, and L. Zhou, *Personalized QoE improvement for networking video service*, IEEE Journal on Selected Areas in Communications, vol. 38, no. 10, pp. 2311–2323, 2020.
13. V. Vasilev, J. Leguay, S. Paris, L. Maggi, and M. Debbah, *Predicting QoE factors with machine learning*, In 2018 IEEE International Conference on Communications (ICC), pp. 1–6, 2018.
14. M.-D. Nguyen, V. Tong, S. Souihi, and A. Mellouk, *Fully-Decentralized Federated Learning for QoE Estimation*, In GLOBECOM 2023–2023 IEEE Global Communications Conference, pp. 5859–5864, 2023.
15. Y. Xu, X. Li, Y. Yang, Z. Lin, L. Wang, and W. Li, *Fedabr: A personalized federated reinforcement learning approach for adaptive video streaming*, In 2023 IFIP Networking Conference (IFIP Networking), pp. 1–9, 2023.
16. P. H. S. Panahi, A. H. Jalilvand, and A. Diyanat, *A new approach for predicting the Quality of Experience in multimedia services using machine learning*, arXiv preprint arXiv:2406.08564, 2024.
17. Darwich, Mahmoud and Bayoumi, Magdy, *Federated Learning for Scalable Video Streaming*, in: Enhancing Video Streaming with AI, Cloud, and Edge Technologies: Optimization Techniques and Frameworks, pp. 61–91, Springer Nature Switzerland, Cham, 2025. doi: 10.1007/978-3-031-84651-9-3
18. Xu, Yeting and Li, Xiang and Yang, Yi and Lin, Zhenjie and Wang, Liming and Li, Wenzhong, *Fedabr: A personalized federated reinforcement learning approach for adaptive video streaming*, 2023 IFIP Networking Conference (IFIP Networking), pp. 1–9, IEEE, 2023.
19. Vo, Phuong L and Nguyen, Nghia T and Luu, Long and Dinh, Canh T and Tran, Nguyen H and Le, Tuan-Anh, *Federated deep reinforcement learning-based bitrate adaptation for dynamic adaptive streaming over HTTP*, Asian Conference on Intelligent Information and Database Systems, pp. 279–290, Springer, 2023.
20. Li, Qinbin and Diao, Yiqun and Chen, Quan and He, Bingsheng, *Federated learning on non-iid data silos: An experimental study*, 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 965–978, IEEE, 2022.