Received: April 22, 2022. Revised: June 27, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Audio-Visual Quality of Experience Prediction Based on ELM Model

Roaa E. Alhassany[1]* Rana Fareed Ghani[1]

[1]*Department of Computer Science, University of Technology, Baghdad, Iraq*
* Corresponding author's Email: roaaeb931@gmail.com

Abstract: Measuring end-user satisfaction, or quality of experience (QoE) became necessary to improve video streaming applications. This measure represents the end-user's degree of satisfaction with the quality of their video conference. This study measures both audio and video QoE, using two types of databases; (UnB-AV database and INRS database) have been used in this work. The UnB-AB database has been used as a target dataset. In this work, several features for audio and video files have been extracted. The extreme learning machine algorithm has been used for predicting the audio-visual QoE, and performance of the proposed model was validated with unseen data. Experiments on the two datasets have shown that the ELM model achieving better prediction accuracy when applied on the UnB-AV database than INRS database. The prediction accuracy by depended on UnB-AV dataset was (0.13) but in depended on INRS dataset was (0.16).

Keywords: Audio features analysis, Extreme learning machine, Multimedia quality, Video streaming, Quality of experience.

1. Introduction

Recently, there has been a significant growth in video conferencing applications, especially with the COVID-19 pandemic, where video conferencing has become a popular method of communicating [1]. Millions of people across several nations are in quarantine, and the internet is the only way that they can communicate, work and attend classes. Applications, such as WebEx, Google Hangouts and Zoom, are being used by millions of people. According to TrustRadius, search impressions for video conferencing applications surged by 500% during the first four months of the COVID-19 outbreak. In 2020, meetings that were conducted over video were 50% more frequent than they were prior to COVID-19. Owl Labs found that 50% of people will not return to employment that does not allow remote work after COVID-19, which emphasises the necessity of investing in the software and technology that are required to host large-scale virtual meetings. According to Upwork, the number of remote employees in the United States will nearly double during the next five years compared to pre-COVID-19 levels. Moreover, by 2025, 36.2 million Americans will be working remotely, which a 16.8 million increase is compared to before the pandemic [2]. The COVID-19 pandemic compelled certain sectors (e.g. the legal and primary healthcare sectors) to convert to digital processes and practices. For example, the number of general practitioners (GPs) who provide video consultations in Norway more than tripled within the first months following the initial lockdown [3].

Therefore, it has become necessary to measure user satisfaction for video conferences because the success of the performance of video conference applications depends on the user satisfaction of the quality of experience (QoE). The QoE is the degree of end-user satisfaction that considers all the elements that impact it [4]. The QoE is a critical measure that network operators and service providers can utilize to assess their performance by considering all elements that influence it [5].

Regarding assessing the video-audio quality, there are the following two concepts to consider: quality of service and QoE. The first is mostly based on well-known network measurements packet loss

rate (PLR), jitter, delay, etc). The latter considers a larger variety of measures to infer user perception of the service quality. As a result, the scientific community has adopted QoE as the best method of assessing the video-streaming performance [6]. Subjective and objective methods are used to assess the QoE. In the subjective approach, many viewers will be judging the video quality. The most common subjective assessment is the mean opinion score (MOS), which is used to assess the video quality and has a scoring range of one to five. Each number represents the degree of user satisfaction with the video quality as follows: one is the worst, two is poor, three is fair, four is good and five is excellent [7]. The objective assessment approach is heavily reliant on reference-based data and uses a mathematical tool [8].

Some research only focuses on the QoE measurement for video streaming and ignores audio streaming. Limited research focuses on the QoE measurement only for voice over internet protocol (VOIP) and does not consider the video streaming. Most previous research only considers the network-based features and ignores the impact of the frequency and time domain features on the QoE. The databases that were used in most previous researches are limited in terms of the source stimuli (SRC) content and degradation diversity. To solve these shortcomings, this paper presents a used machine-learning method to predict the QoE based on audio-visual information using a database with content and degradation diversity that considers the network-based features and frequency and time domains.

This paper is organized as follows: In Section 2 we discuss the literature review. Section 3, in this section explains the methodology that has been followed to build the proposed system. Section 4 presents the results of discussion while the conclusion is presented in Section 5.

1.1 Contributions

The contribution of this paper is three-fold as follows:

- Use machine learning to predict the QoE based on visual and audio information.
- Extract many features for audio and video that have a high impact on the QoE.
- The proposed model has been trained using a dataset with SRC content and degradation diversity.

2. Literature review

Several studies that measure the QoE have been published and can be categorised into video and audio QoE measuring. Bao et al. [9] evaluated the video QoE using a fuzzy clustering heuristic algorithm. They used the server side to save some information and quality of service (QoS) parameters in a large database. The heuristic rules model used the data that was saved in the database to predict the user scores. This method is called a fuzzy clustering analysis and provides a service QoE that is sent to a customer. Mohamed Alreshodi [10] presented a fuzzy inference system (FIS) model to predict the video quality. The authors investigated the effects of QoS parameters on QoE for a variety of video content types and assessed video quality from the MOS perspective. When the suggested system was compared with the regression-based system, the FIS model provided a higher accuracy. Ghani et al. [7] presented a no-reference technique for assessing the video QoE using machine-learning algorithms (AdaBoost, C45, Random Forest, multilayer perception, artificial neural network (ANN)), comparing its performance and choosing the best algorithm that balances time and precision. They used bit rate with pixel mode features to predict the MOS. The best real-time and accuracy performance was provided by the AdaBoost decision tree. Sufiuh et al.[11] presented an ANN algorithm to predict the video QoE. They extracted seven features and utilized them as inputs for training data. The features that were extracted were as follows: temporal perceptual information (SPI), freezing, blurring, luminance, an average of luminance difference and blocking. The result provides the best correlation between the predicted and measured QoE. ZhiGuo et al. [8] comprehensively analysed the effects of QoS on the QoE instead of analysing each parameter separately using experimental methods and an association test technique. The authors proposed an algorithm to combine the impact of QoS with the temporal or spatial features on QoE. Finally, they applied several machine-learning regression algorithms (K-nearest neighbours (KNN), support vector machine (SVM), Regression tree, Bagging, ANN) for different QoS degradations, echoes and noises in a diverse network environment to predict the non-intrusive voice quality. The result was accurate for the VOIP QoE evaluations when utilising QoS parameters. Charonyktakis et al. [12] proposed the modular user-centric algorithm MLQoE. The correlation between the QoE and network QoS metrics for VOIP services is based on supervised learning. The

approach is modular in that it trains various supervised learning models based on gaussian naive bayes (GNB), decision tree (DT), support vector regression (SVR), multilayer perceptron (MLP) and artificial neural network (ANN), then chooses the most accurate model after cross-validation. Compared to other existing machine-learning models, MLQoE can accurately predict the QoE score. Demirbilek et. [13] using the quality of an audio-visual dataset obtained from Institute National de la Recherché Scientific (INRS), they developed several no-referenced Machine Learning (ML), to compare between accuracy performance for genetic programming, ensemble decision tree, deep learning. When the INRS dataset was compared to other datasets, it was discovered that random forests outperformed other prediction models in terms of accuracy. In terms of RMSE and Pearson correlation measures, decision trees outperform both deep learning and genetic programming, they attain RMSE results for all models in the range of (0.340) to (0.469).

We have summarised the research issues of the video QoE measuring as follows:

- Ignore the audio QoE measurement.
- The dataset that was used was limited in terms of the SRC content and degradation diversity.
- Use a small dataset size.
- Machine-learning algorithms that have been used suffered from slow data training process.
- AdaBoost in [7] requires enough data and many iterations to achieve acceptable accuracy, which may result in increased time complexity. It is not appropriate for predictions that require high speed.

Most previous research used datasets depending on the ACR method, and the ACR method is not accurate.

The following summarises the research issues with audio QoE measurements:

- Ignore audio QoE measurements.
- The frequency and time domain features for audio were not considered.
- Use a small dataset size and small number of features.
- The dataset that was used was limited in terms of the SRC content and degradation diversity.
- Machine-learning algorithms that have been used suffered from slow data training.

3. Methodology

The methodology includes the following six sections: datasets, features extraction, features selection, normalisation, unnormalisation and model steps. Fig. 1 shows the proposed system phases.

3.1 Datasets

The following two types of datasets were used in this work: UnB-AV and INRS datasets. The two datasets are explained as follows:

3.1.1. The UnB- AV database (target)

This consists of three datasets of audio-visual databases, each of which was utilised in a different subjective experiment. The three datasets were created using (140) high-definition video sequences (with associated audio) as the source. They were divided into the following three groups: (60) video sequence for experiment one; (40) video sequence for experiment two; and (40) video sequence for experiment three. Additionally, it included 2,320 test sequences with audio and video degradations, coupled with the psychophysical experiments that assessed the audio-visual quality of a series of video sequences in all three experiments. For the subjective studies, each experiment used the immersive technique. In the first experiment, although the video component was impaired by visual artifacts (PLR, frame freezing and video coding), it did not degrade the audio component. For the second experiment, when the signal artifacts (clipping, chop, background noises and echoes)

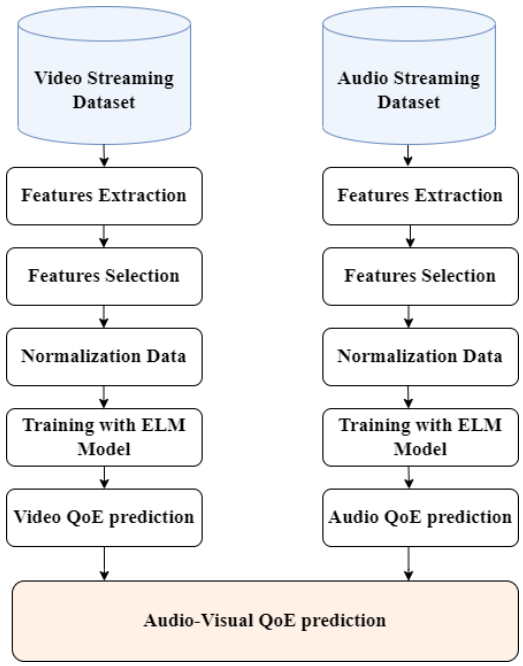


Figure. 1 The proposed system phases

were applied to the audio component, the video component remained unaffected. Lastly, both the video and audio components were impaired regarding the same kinds of degradations that were utilised for the previous two experiments. The subjects were asked to assess the overall audio-visual quality in all three experiments. It used the following two statistical analysis methods: mean quality score (MQS) and mean [14, 15].

3.1.2. INRS audio-visual quality database

In this dataset, the audio and video quality were affected by some degradations of the PLR, quantization parameter and video frame rates. It contained 160 audio-visual files that were extracted from a single source video. The H.264/AVC was used to encode the source video. The H.AMR-WB by Gstream (an open-source (O/S) framework) was used to create 32 GP videos as a source at various quality levels, with a quantisation parameter (QP) of 10 fps, 15 fps, 20 fps and 25 fps and frame rate (FR) of 10 fps, 15 fps, 20 fps and 25 fps. The network emulator was used to generate the PLR to obtain more realistic outcomes. However, it was only enabled after the first second had been transmitted. The audio-visual file name was included in the dataset, as well as the computed MOS, various parameters that were collected from broadcasted videos and the resolution of the videos (720p) [16].

3.2 Feature extraction

3.2.1. Video feature extraction

The number of features has been extracted to enhance the prediction model accuracy. The extracted features have a significant linkage to the human visual system (HVS). The extracted features are presented as follows:

3.2.1.1. Blocking feature

Kirsch compass masks were used to extract blocking features by rolling one mask through the following eight typical compass orientations: W, SW, N, NW, S, SE, NE and E. The Krisch mask directions are shown in Eq. (1) [17].

$$G_1=\begin{bmatrix}-3 & -3 & 5\\ -3 & 0 & 5\\ -3 & -3 & 5\end{bmatrix} G_2=\begin{bmatrix}-3 & 5 & 5\\ -3 & 0 & 5\\ -3 & -3 & -3\end{bmatrix}$$
$$G_3=\begin{bmatrix}5 & 5 & 5\\ -3 & 0 & -3\\ -3 & -3 & -3\end{bmatrix} G_4=\begin{bmatrix}5 & 5 & -3\\ 5 & 0 & -3\\ -3 & -3 & -3\end{bmatrix}$$
$$G_5=\begin{bmatrix}5 & -3 & -3\\ 5 & 0 & -3\\ 5 & -3 & -3\end{bmatrix} G_6=\begin{bmatrix}-3 & -3 & -3\\ 5 & 0 & -3\\ 5 & 5 & -3\end{bmatrix}$$
$$G_7=\begin{bmatrix}-3 & -3 & -3\\ -3 & 0 & -3\\ 5 & 5 & 5\end{bmatrix} G_8=\begin{bmatrix}-3 & -3 & 5\\ -3 & 0 & 5\\ -3 & 5 & 5\end{bmatrix} \quad (1)$$

0	1	0
1	-4	1
0	1	0

Figure. 2 Laplacian kernel

3.2.1.2. Blurring feature

This feature is very important for NR features. The blur feature is extracted by the implemented Laplacian operator. The Laplacian draws attention to parts of an image that have a lot of intensity variations. The image is convolved with a 3×3 Laplacian operator. The variance is computed using the Laplacian kernel for each frame, after which the blur average is calculated for the video frames, as shown in Fig. 2 [18].

3.2.1.3. Natural senses statistics (NSS)

The use of an (NSS) model when choosing the perceptual features that can provide the satisfying result regarding the problem of video quality evaluation is a significant research direction. The features that have been extracted are as follows: (N_H Shape, N_H Variance, N_Shape, N_V Shape, N_Shape and N_Variance). The NSS was extracted using the Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) method. It calculates the contrasting normalised coefficients for the asymmetric generalised Gaussian distribution [19].

3.2.1.4. Average bit frame

The coding parameters were used to extract the average bit frame, which increases the accuracy of the prediction performance. The relationship between the bit and frame rates is strong, has a major influential and is linked to the end QoE. Eq. (2) was used to calculate the average bit rate, and it represents the average number of bits that were utilised to represent a single pixel. The height and width of a video indicate its resolution [7].

$$\text{AVGBF} = \text{BR} / \text{Height} \times \text{Width} \times \text{Fr} \quad (2)$$

3.2.2. Audio feature extraction

To obtain a more precise regression of audio, the number of features must be extracted. To extract

valuable features from prepared audio files, we must use the Librosa library, which is a python package that is usually used for audio and music signal processing analyses. It has numerous functions for feature extractions, filters and spectral, temporal segmentation [20]. Table 1 describes the features that were extracted.

Table 1. The features extracted from audio

Categories of Features	Features	Descriptions
Time-Domain	Zero-Crossing Rate (ZCR)	It emphasizes how often the signal changes from negative to positive and vice versa [21].
	Root Mean Square (RMS)	RMS of a signal represents the signal's power content [22].
	Tempogram	Pulse intensity over time for a particular time lag 1 or BPM value τ [23].
	Fourier-tempogram	A tempogram is a time-tempo representation that encodes the local tempo of a music signal over time [23].
Frequency Domain	Mel-Frequency Cepstral Coefficients (MFCC)	It is one of the most sophisticated technologies and is based on the fact that the crucial bandwidths of the human ear vary in frequency. The Mel-frequency scale, which is a linear frequency space below 1000 Hz and a logarithmic space above 1000 Hz, is used to show this information [24]
	Spectral Centroid (SC)	SC (also known as brightness) represents the focal point in the spectral power distribution of a signal in a sample frame [25].
	Mel Spectrogram	It stimulates the biological auditory systems of humans by producing a temporal frequency representation of sound [26].
	Tonal Centroid features (Tonnetz)	Changes in the harmonic content of musical audio signals, such as chord boundaries in polyphonic audio recording, can be detected using these features [26].
	Spectral Bandwidth	It is the second-order statistical value that distinguishes low-bandwidth sounds from high-frequency sounds. It is commonly utilized in music classification and sound identification in the environment [27].
	Spectral Contrast	The decibel difference between peaks and troughs in the spectrum is known as a spectral contrast [28].
	Spectral Roll-off	It is characteristically defined as the frequency at which 95 per cent of spectral energy in a signal is collected [29].
	Chroma Energy Normalized (CENS)	It is widely utilized in the field of musical signal processing. Because Chroma features retain melodic and harmonic qualities of music and are resistant to changes in instrumentation and timber, they are useful in audio matching and retrieval applications [26].
	Chroma(stft)	It is a well-known technique for analysing a signal's frequency distribution [30].
	Poly-features	It is used to calculate the coefficients of fitting an nth-order polynomial to a spectrogram's columns [31].
Time-Frequency Domain	Chroma Constant-Q Transform (CQT)	Converts a data series from the time domain to the 4frequency domain. It has something to do with the Fourier transform [32].

3.3 Features selection

In this stage, we used the information gain method to select the best features that have a high impact on the audio-visual QoE. The features with the highest impact on the video streaming were as follows: average bit frame, PLR, QP, freezing frame, blocking and blur. The features that provided the best and highest impact and correlation on the audio QoE were as follows: chop, clipping, echoes, noises, ZCR, MFCC, SC, tempogram, Mel spectrogram, RMS, Tonnetz,spectral bandwidth, spectral contrast, spectral roll-off, CQT, CENS, Chroma (stft) and poly-features.

3.4 Normalisation and unnormalisation

All input and output variables were transformed into ranges (zero to one) using the MinMax scaler method.

The following equations were used for normalisation and unnormalisation.

$$Xn_i = (X_i\text{-min}(X_i)/ (\max(X_i)\text{-min}(X_i)) \quad i=1,2\ldots.n \tag{3}$$

$$Yn_p= (Y_p\text{-min}(Y_p)/ (\max(Y_p)\text{-min}(Y_p)) \quad ,p=1,2\ldots2 \tag{4}$$

Where:
Xn_i: i^{th} normalized input value in the dataset.
X_i : i^{th} input value in the dataset.
$\text{Min}(X_i)$: minimum input value in the dataset.
$\text{Max}(X_i)$: maximum input value in the dataset.
Yn_p: i^{th} normalized output value in the dataset.
Y_P : i^{th} normalized output value in the dataset.
$\text{Min}(Y_p)$: minimum normalized output value in the dataset.
$\text{Max}(Y_p)$: maximum normalized output value in the dataset.

The following equation is used to convert data back into unnormalize units:

$$Xun_i=Xn_i\times(\max(X_i)\text{-}(\min(X_i) +\min(X_i), \quad i=1,2,3...n \tag{5}$$

$$Yunp=Ynp\times (\max(Yp) - \min(Yp) +\min(Yp), \quad p=1,2\ldots n \tag{6}$$

Where:
Xun_i: i^{th} unnormalized input value in the dataset.
Xn_i : i^{th} input value in the dataset.
$\text{Min}(X_i)$: minimum input value in the dataset.
$\text{Max}(X_i)$: maximum input value in the dataset.
Yun_p: i^{th} unnormalized output value in the dataset.
Yn_P : i^{th} unnormalized output value in the dataset.
$\text{Min}(Y_p)$: minimum unnormalized output value in the dataset.
$\text{Max}(Y_p)$: maximum unnormalized output value in the dataset.

The minimum and maximum values in the dataset were selected using the min (.) and max (.) operators in Eqs. (3) to (6).

3.5 Proposed model overview

The extreme learning machine (ELM) is a single hidden layer feed-forward neural network (SLFN) that chooses hidden nodes at random and calculates the output weights of SLFNs analytically [33]. As a result, it only requires one iteration [34] and does not need to train in an iterative way like traditional neural networks. The hidden layer's output weights are calculated by taking the generalised inverse of its output. This procedure enhances the network construction of the ELM [35]. In general, this algorithm provides an excellent generalisation performance at a fast learning rate. The traditional feed-forward neural network learning methods are much slower than ELMs. Typically, on average, ELMs will achieve the least output weight norms and training errors [36]. Training data is used to create a prediction when using ML to predict video QoE assessments. This prediction should be able to generalise effectively to fresh data with no ground truth [37].

3.5.1. ELM steps

ELM algorithm steps are presented as follows:

Input: Set of training samples se$t\ \{x_i, t_i\}_{i=1}^{N} \subset R^n \times R^m$, set of testing samples $\{y_i\}_{i=1}^{M} \subset R^n$, L is the number of hidden layer nodes and g(.) is the activation function.

Step 1: Th**e** output matrix H in the hidden layer is computed using the following equation:

$$(c_1,..,c_L; b_1..b_L; x_1..x_n) = \begin{bmatrix} g(c_1,b_1,x_1) & g(c_L,b_L,x_1) \\ \vdots & \vdots \\ g(c_1,b_1,x_N) & g(c_L,b_L,x_N) \end{bmatrix}_{N\times L} \tag{7}$$

Where (c_i,b_i), i=1,2,3,……,L are hidden node parameters that are produced at random, where c_i represents the input weight of the ith hidden layer node, and bi represents the deviation of the ith hidden layer node.

Step 2: To compute the output weight matrix β in the hidden layer, Eq. (8) was used.

$$\beta = H^{+} T \quad (8)$$

Where β is the output weight, H^{+} is the generalized Moore-Penrose inverse matrix of hidden node output matrix H, and T is the target output.

Output: Weight matrix b output.

Supervised techniques were used by the proposed model as a regression system to predict the video QoE using the training dataset. We trained the proposed model with two types of datasets to improve the accuracy prediction. As the INRS dataset had one SRC content and limited degradation diversity, to achieve a more accurate prediction, we used the UnB-AV dataset. In the UnB-AV dataset, as experiment one only contained audio information, we trained this information to predict the audio QoE. Experiment two contained video information that had been trained to predict the QoE. Finally, to predict the audio-visual QoE, we used experiment three to train the proposed model. The proposed model (ELM) consists of the following three layers: layer one used many features as inputs for the dataset; the second layer is the hidden layer that includes some neurons and is equal to the number of inputs for each dataset; and the last layer contains one neuron that represents the output and reflects the predictions for audio, video or audio-visual QoE. The rectified linear unit (ReLu) was used as an activation function. This work was achieved using Python language, version 3.8.

4. Results and discussion

The validation of the proposed prediction model used the unseen data (testing data). The actual values of the QoE were compared with the predicted values of the QoE. To evaluate the proposed model's accuracy regarding predicting QoE for audio, video and audio-visual information using the UnB-AV dataset, the following three types of validation metrics were used: root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE). As shown in Table 2, although the RMSE score was 0.18 in the UnB-AV dataset for predicting audio QoE (experiment 2), its prediction error rate was less than for the MSE and MAE, which scored 0.03 and 0.13, respectively. After merging the audio and video information to predict the audio-visual QoE and evaluating the accuracy, we found that the model achieved a high accuracy and scored an RMSE of 0.13 when predicting with MQS.

To evaluate the ELM model performance, by using the INRS dataset that was used for predictions with audio-visual QoE, we compared our results with the results for latest methods from other researchers. In [13], as mention in literature review section, authors they used their proposed system to predict the QoE for audio-visual and obtained an RMSE of 0.34–0.46 to evaluate the model accuracy. However, the proposed model scored a lowest error rate and better RMSE value of 0.13 for predictions with audio QoE. When we merged the audio and video information, we obtained the lowest accuracy regarding predicting the QoE for video only, with a RMSE score of 0.13. In [7], they implemented RMSE to measure the model's performance and obtained a score of 0.07. Table 3 shows the different accuracy and correlation metrics from different researchers.

Table 2. Accuracy results for predicting QoE with the proposed system for the UnV-AV dataset

Types of Multimedia	RMSE	MAE	MSE
Video	0.11	0.08	0.012
Audio	0.18	0.13	0.03
Audio-Visual	0.13	0.10	0.01

Table 3. Performances of the ELM model in comparison with other researcher

Previous Research	Types of Multimedia	RMSE
Demirbilek [13]	Audiovisual	0.340–0.469
Osama [7]	Audio-Visual	0.072
The Proposed Model (ELM)	Video	0.11
	Audio-Visual	0.13

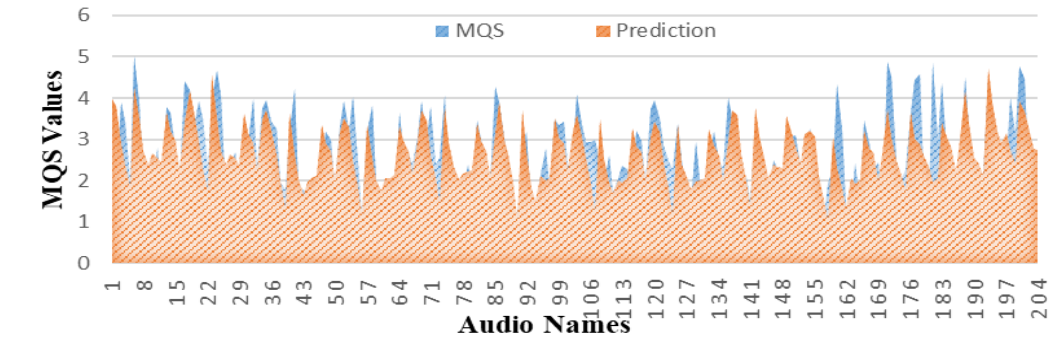
From the results for the proposed model on UnB-AV and INRS datasets, we can conclude the following:

The difference between the actual values for QoE and predictions of QoE was insignificant. As shown in Fig. 3, the prediction for QoE is represented by the orange colour, and the blue colour represents the measured QoE. This refers to the proposed model that was appropriate for predictions with audio, video and audio-visual QoE. Table 4, and Table 5 shows a sample of a comparison data between the actual QoE values and predicting QoE values.

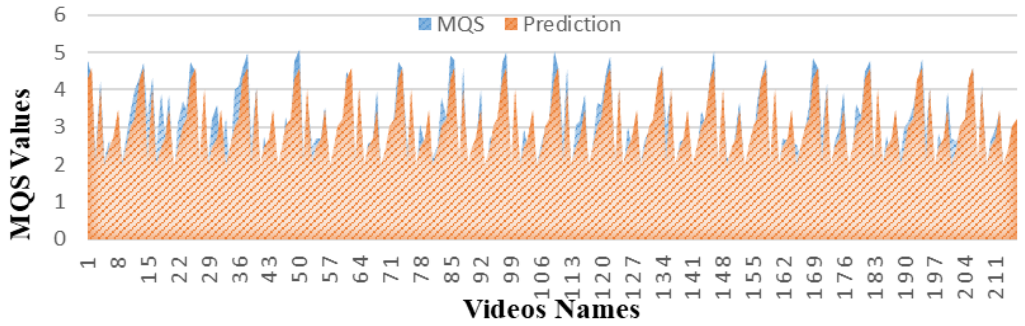
- The extracted features improved the prediction model accuracy.
- Most features that were extracted for audio and video had a high impact on the QoE. Previous research proved that the features that were extracted for video streaming had a high impact on the video QoE. However, in this paper, we proved that the frequency and time domain features that we extracted for audio had a high

impact on the QoE. As shown in Fig. 4, all time and frequency domain affected the QoE. Fig. 5 and Table 5 shows the correlation and impact for the features that we extracted for audio on the QoE. For example, the QoE becomes excellent (five value) in the high value tempogram and has a bad value when the tempogram value starts to decrease.

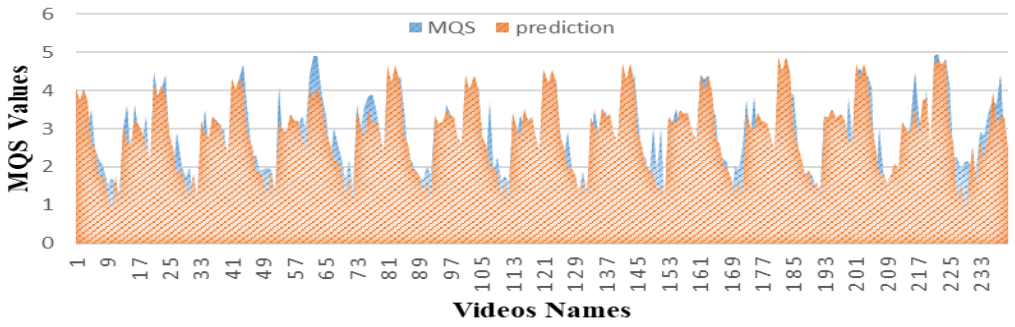
- As it did not take much time to run the ELM model, this model was very appropriate form Predicting audio-visual QoE.



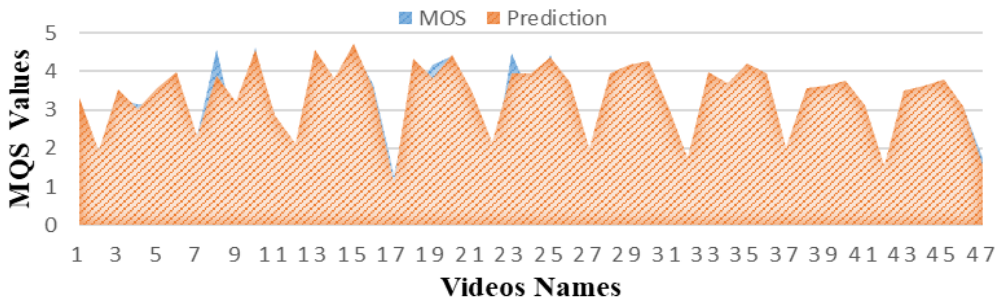
(a)



(b)



(c)



(d)

Figure. 3 (a) Comparison between actual value and the prediction value for audio information (UnV-AV dataset), (b) Comparison between actual value and the prediction value for video information (UnV-AV dataset), (c) Comparison between actual value and the prediction value for audio-visual information (UnV-AV dataset), and (d) Comparison between actual value and the prediction value for audio-visual information (INRS dataset)

Table 4. A comparison between actual QoE values and predicting QoE values for UnB-AV database

Actual Videos QoE Values	Predicting Videos QoE Values	Actual Audio QoE Values	Predicting Audio QoE Values	Actual Audio-Visual QoE Values	Predicting Audio-Visual QoE Values
4.75852	4.27645	3.9	3.96522	3.75	4.01841
4.35375	4.57145	2.02667	3.81961	3.38	3.74369
2.16065	2.04329	3.88	3.88113	3.81333	4.01841
4.21407	4.00417	3.43333	2.39872	3	3.74369
2.13355	1.98871	1.97	1.86384	3.47403	2.65192
2.63561	2.45623	5	4.25813	2.4675	2.49791
1.4873	2.70355	3.86	3.52104	2.2	1.69059
3.25581	3.46851	2.02333	2.71415	2.06	1.8228
2.07622	1.95521	1.86294	1.79466	1.52167	1.32178
2.63257	2.42244	1.678	2.64895	1.715	0.881207
3.46775	3.02543	2.8	2.55977	1.55	1.77569
3.9785	3.22963	1	1.22104	1.31333	1.03931

Table 5. A comparison between actual QoE values and predicting QoE values for INRS database

Actual Videos QoE values	Predicting Videos QoE Values	Actual Audio QoE values	Predicting Audio QoE Values	Actual Audio-Visual QoE Values	Predicting Audio-Visual QoE Values
2.791	2.69796	3.55767	3.20749	2.791	3.31601
1.46306	1.32924	3.19195	2.791	1.46306	1.93709
3.26803	3.17725	1.58958	1.46306	3.26803	3.54332
3.15656	3.01367	3.57578	3.26803	3.15656	3.01141
3.19536	3.21605	3.23443	3.15656	3.19536	3.54916
3.08753	2.82422	3.55767	3.19536	3.08753	3.97399
2.29991	2.21667	3.13231	3.08753	2.29991	2.29845
4.56546	3.58408	1.67448	2.29991	4.56546	3.87346
2.38523	3.16989	3.438	4.56546	2.38523	3.2071
4.61638	3.68694	3.15343	3.18523	4.61638	4.57469
2.01926	2.747	3.55767	4.61638	2.01926	2.85533
1.31538	0.494684	2.6475	2.41926	1.31538	2.11228

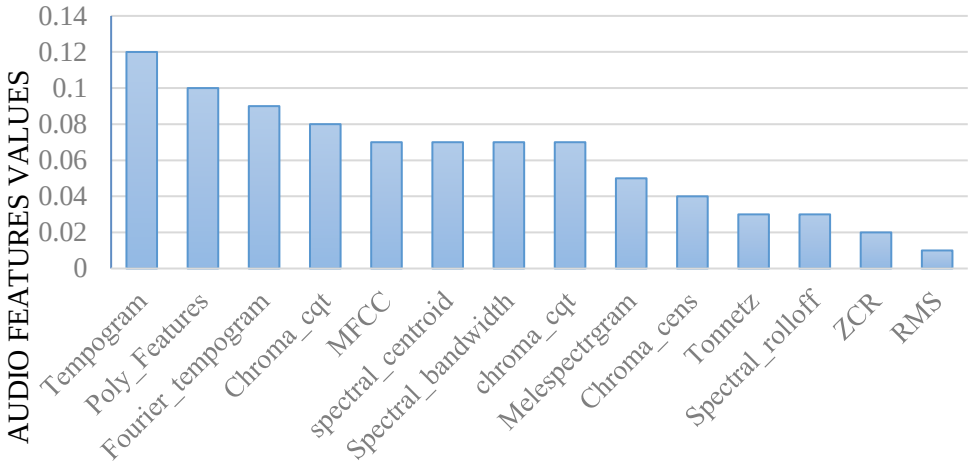


Figure. 4 The Correlation between time and frequency domain with the audio QoE

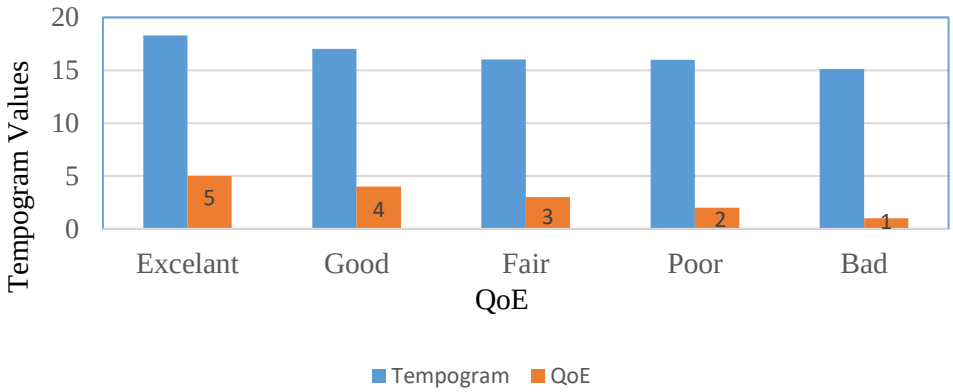


Figure. 5 The QoE depend on the tempogram feature

Table 6. Shows the sample of data for the time and frequency domain impacts on the QoE

Tempogram	Chroma_ Cqt	Spectral_ Centroid	Spectral_ Bandwidth	Fourier_ Tempogram	Poly_Features	RMS	QoE
18.01216	0.59107	2231.431	2258.473	0.270678	0.27532	0.062137	4.87
16.45556	0.555938	1778.193	2234.007	0.273136	0.268275	0.037583	2.68
14.24353	0.619521	3196.019	3655.905	0.289347	0.279878	0.062137	2.2075
13.76668	0.631257	3726.78	4186.354	0.300147	0.280906	0.062137	2.085982
13.32156	0.642852	4653.118	4995.482	0.309395	0.281661	0.062137	2.31
17.87725	0.586376	2262.033	2285.511	0.271771	0.273672	0.063628	3.275
17.82678	0.593706	2275.049	2310.799	0.277721	0.274342	0.063263	3.285
17.71329	0.599856	2320.073	2379.652	0.280681	0.275407	0.063882	2.558075
17.58824	0.608973	2395.336	2472.675	0.288025	0.276311	0.065571	1.39
16.38276	0.6057	2333.67	2545.891	0.281416	0.279026	0.061589	2.46
15.00344	0.598923	2460.507	3108.749	0.277051	0.27586	0.057946	1.220
16.08505	0.566087	2258.991	2560.545	0.295808	0.266284	0.051955	2.5375
17.42992	0.595705	2249.946	2313.242	0.268995	0.275791	0.062149	3.476667
17.16071	0.594891	2281.125	2380.3	0.235696	0.275939	0.059616	3.525
17.10775	0.594919	2265.946	2370.696	0.266979	0.275835	0.061742	4.52
16.46979	0.604454	2356.154	2558.504	0.207616	0.276899	0.05609	2.6575
14.68198	0.612263	2924.454	3342.863	0.281545	0.279042	0.062137	1.99
20.04716	0.53414	2805.258	3569.874	0.243772	0.256659	0.049524	4.24
17.55556	0.593116	2454.545	3399.705	0.26444	0.268291	0.039527	3

5. Conclusions

This paper proposed a method for measuring audio-visual information QoE based on machine learning algorithms. The study used two types of datasets: training and testing datasets. The UnV-AV dataset was used as a target dataset because it contains diverse audio-visual material (e.g. TV commercials, music, documentaries, cartoons, sports and interviews) and some degradations. Several features that have a high impact on the audio-visual QoE were extracted to increase the prediction accuracy. This study consisted of three experiments: predicting audio, predicting video and predicting merged audio-visual. The results indicate that audio degradations have a high impact on the QoE. This study proved that the time and frequency domain features significantly impact the audio QoE. Table 6 shows the sample of data for the time and frequency domain impacts on the QoE.

Extreme learning machine (ELM) model has been used to predict QoE, and it is validated using testing data (unseen data). We can conclude that the ELM model is highly suited to predicting audio-visual information because it produces minor training errors and provides generalized execution because of the algorithm structure. The result

indicates that the ELM model was appropriate for predicting audio-visual QoE at high speed; it does not have to learn iteratively and it needs a single iteration process. Experiments on the two datasets revealed that the ELM model performed better on the UnB-AV database than the INRS database in terms of prediction accuracy. The prediction accuracy while using the UnB-AV dataset was (0.13), whereas when using the INRS dataset. It was (0.16). In this work the shortcomings in the previous researches have been solved by predicting not only video QoE or only audio QoE, but has been predicted with both audio and video information. The audio-video QoE prediction in this work depend not only on the network features but also on the pixel mode features for video and Frequency-time domain for audio.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

Conceptualization, by 2nd author. ; methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] J. Lin, P. Liu, Y. Zheng, W. Deng, and M. Zeng, "Real-Time Masked Face Revealing for Video Conference", *IEEE International Conference on Multimedia and Expo*, Vol. 1, 2021.

[2] "18 video conferencing statistics for 2021.", https://blog.webex.com/video-conferencing/18-video-conferencing-statistics-for-2021/ (accessed May 27, 2021).

[3] E. Øie, K. Koniuch, N. Cieplińska, and K. D. Moor, "Factors influencing QoE of video consultations", *International Conference on Quality of Multimedia Experience*, Vol. 2, 2021.

[4] M. Caro and M. D. Cano, "On the Identification and Prediction of Stalling Events to Improve QoE in Video Streaming", *Electronics*, Vol. 10, No. 6, pp. 1-753, 2021.

[5] L. Du, L. Zhuo, J. Li, J. Zhang, X. Li, and H. Zhang, "Video quality of experience metric for dynamic adaptive streaming services using DASH standard and deep spatial-temporal representation of video", *Applied Sciences*, Vol. 10, No. 5, p. 1793, 2020.

[6] T. Abar, A. B. Letaifa, and S. E. Asmi, "Quality of experience prediction model for video streaming in SDN networks", *International Journal of Wireless and Mobile Computing*, Vol. 18, No. 1, pp. 59-70, 2020.

[7] R. Ghani and O. Shalal, "Objective video streaming QoE measurement based on prediction mode", *Computer Science and Electronic Engineering*, pp. 201-206, 2017.

[8] Z. Hu, H. Yan, T. Yan, H. Geng, and G. Liu, "Evaluating QoE in VoIP networks with QoS mapping and machine learning algorithms", *Neurocomputing*, Vol. 386, pp. 63-83, 2020.

[9] Y. Bao, W. Lei, W. Zhang, and Y. Zhan, "QoE collaborative evaluation method based on fuzzy clustering heuristic algorithm", *Springerplus*, Vol. 5, No. 1, pp. 1-29, 2016.

[10] M. Alreshoodi, "Prediction of Quality of Experience for Video Streaming Using Raw QoS Parameters", *University of Essexy*, 2016.

[11] A. Ajrash, R. Ghani, and L. A. Jobouri, "ANN based measurement for no-reference video quality of experience metric", *Computer Science and Electronic Engineering*, 2019.

[12] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, "On user-centric modular QoE prediction for voip based on machine-learning algorithms", *IEEE Transactions on Mobile Computing*, Vol. 15, No. 6, pp. 1443-1456, 2016.

[13] E. Demirbilek and J. C. Grégoire, "Perceived audiovisual quality modelling based on decison trees, genetic programming and neural networks", *arXiv Preprint arXiv:1801.05889*, pp. 1-14, 2017.

[14] H. Martinez, A. Hines, and M. Farias, "Perceptual quality of audio-visual content with common video and audio degradations", *Applied Sciences*, Vol. 11, No. 13, p. 5813, 2021.

[15] H. Martinez, A. Hines, and M. Farias, "UnB-AV: An Audio-Visual Database for Multimedia Quality Research", *IEEE Access*, Vol. 8, pp. 56641-56649, 2020.

[16] E. Demirbilek and J. Grégoire, "INRS audiovisual quality dataset", In: *Proc. of ACM International Conference on Multimedia*, pp. 167-171, 2016.

[17] R. Ranjbarzadeh, S. Saadi, and A. Amirabadi, "LNPSS: SAR image despeckling based on local and non-local features using patch shape selection and edges linking", *Measurement*, Vol. 164, p. 107989, 2020.

[18] R. Bansal, G. Raj, and T. Choudhury, "Blur image detection using Laplacian operator and

Open-CV", In: *Proc. of International Conference System Modeling & Advancement in Research Trends*, pp. 63-67, 2017.
[19] A. Sadiq, I. Nizami, S. Anwar, and M. Majid, "Blind image quality assessment using natural scene statistics of stationary wavelet transform", *Optik*, Vol. 205, p. 164189, 2020.
[20] Z. Mushtaq and S. Su, "Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images", *Symmetry*, Vol. 12, No. 11, pp. 1-34, 2020.
[21] K. Racharla, V. Kumar, C. Jayant, A. Khairkar, and P. Harish, "Predominant musical instrument classification based on spectral features", In: *Proc. of International Conference on Signal Processing and Integrated Networks*, pp. 617-622, 2020.
[22] S. A. Agha, H. Saleh, and R. Ghani, "Analyze Features Extraction for Audio Signal with Six Emotions Expressions", *Int. J. Eng. Adv. Technol*, No. 6, pp. 2249-8958, 2015.
[23] M. Tian, G. Fazekas, D. Black, and M. Sandler, "On the use of the tempogram to describe audio content and its application to Music structural segmentation", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 419-423, 2015.
[24] T. Liu, D. Yan, R. Wang, N. Yan, and G. Chen, "Identification of fake stereo audio using svm and cnn", *Information*, Vol. 12, No. 7, p. 263, 2021.
[25] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomedical Signal Processing and Control*, Vol. 59, p. 101894, 2020.
[26] D. Ghosal and M. H. Kolekar, "Music genre recognition using deep neural networks and transfer learning", *Interspeech*, pp. 2087-2091, 2018.
[27] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods", *Applied Acoustics*, Vol. 158, p. 107020, 2020.
[28] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification", *Applied Acoustics*, Vol. 158, p. 107050, 2020.
[29] M. Rahman, B. Sultana, and A. Khatun, "Classification of Bangla Alphabets Phoneme based on Audio Features using MLPC & SVM", In: *Proc. of International Conference on Automation, Control and Mechatronics for Industry 4.0*, pp. 1-5, 2021.
[30] S. Rajesh and N. J. Nalini, "Musical instrument emotion recognition using deep recurrent neural network", *Procedia Computer Science*, Vol. 167, pp. 16-25, 2020.
[31] A. Shoiynbek, K. Kozhakhmet, N. Sultanova, and R. Zhumaliyeva, "The robust spectral audio features for speech emotion recognition", In: *Appl. Math*, Vol. 13, No. 5, pp. 867-870, 2019.
[32] K. W. Cheuk, K. Agres, and D. Herremans, "The Impact of Audio Input Representations on Neural Network based Music Transcription", In: *Proc. of International Joint Conference on Neural Networks*, pp. 1-6, 2020.
[33] R. Choudhary and S. Shukla, "A clustering based ensemble of weighted kernelized extreme learning machine for class imbalance learning", *Expert Syst. Appl*, Vol. 164, No. September 2020, p. 114041, 2021.
[34] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications", *Neurocomputing*, Vol. 70, No. 1-3, pp. 489-501, 2006.
[35] Z. Pan, Z. Meng, Z. Chen, W. Gao, and Y. Shi, "A two-stage method based on extreme learning machine for predicting the remaining useful life of rolling-element bearings", *Mech. Syst. Signal Process*, Vol. 144, p. 106899, 2020.
[36] A. Jahromi, S. Hashemi, K. R. Choo, and R. Parizi, "An improved two-hidden-layer extreme learning machine for malware hunting", *Computers & Security*, Vol. 89, p. 101655, 2020.
[37] M. Giannopoulos, G. Tsagkatakis, S. Blasi, and E. Izquierdo, "Convolutional neural networks for video quality assessment", *arXiv Preprint arXiv:1809.10117*, 2018.