Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

# A Critical Analysis of Deep Learning Methods for Video QoE Prediction

***Salwa Aqeel Mahdi[a], Huda Abdulaali Abdulbaqi[b], Hazeem B. Taher[c]***

*[a]Mustansiriyah University, College of Science, Computer Science Department, Baghdad Iraq.Email: gosalwyahoo.com@uomustansiriyah.edu.iq*
*[b]Mustansiriyah University, College of Science, Computer Science Department, Baghdad Iraq.Email: huda.it@uomustansiriyah.edu.iq*
*[c]Ministry of Higher Education and Scientific. Email: mailto:hazecomp792004@gmail.com, hazeem.sci@utq.edu.iq*

ARTICLE INFO

*Article history:*
Received: 23/01/2025
Rrevised form: 03/04/2025
Accepted : 22/04/2025
Available online: 30/06/2025

*Keywords:*

QoE, Deep Learning, CNN, LSTM, Classification, RegressionEach keyword to start on a new line

ABSTRACT

Multimedia video applications significantly impact video quality prediction, widely regarded as one of the most challenging problems. The Quality of Experience (QoE) prediction of the video mimics the satisfaction of the content of the video as humans perceive it. Machine learning and deep learning models have applied numerous methods to obtain QoE predictions. Some of these methods are full reference or reduced reference (half reference); others are no reference. In this paper, we attempt to explore, evaluate, and analyze the different scenarios and models related to QoE predictions for videos using deep learning. We have conducted a comprehensive examination to address the limitations of the existing models. Moreover, we suggest a new framework to overcome the limitations of the existing models.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.22176

## 1. Introduction

Nowadays, with a high demand for multimedia, particularly video content, people are more likely to communicate via wireless channels. Universally, video content dominates Internet traffic data with about 92% of total data used across the Internet [1], as illustrated in figure 1.Wireless channels limited bandwidth necessitates the compression of videos for streaming. This leads to video degradation. In addition, videos always get distortion because of the noise. These reasons may lead the user to get a bad-quality video. As a result, the user will give up the provider's services or not continue watching the video. The providers of services have to check the quality of the videos. Multimedia streaming services (videos) use QoE as a metric to gauge user satisfaction. To improve video viewing, human involvement is needed to give a subjective assessment of video quality. However, human assessment is costly and time-consuming. Furthermore, people's psychological state and surroundings influence their judgement when evaluating the quality of videos. All these difficulties led to the objective prediction of the video.

The quality of video experience has been widely investigated in multimedia communication studies, although video prediction is regarded as a difficult task. The quality of the video depends on many factors called influencer factors (IF). We can divide these factors into three groups: system factors, context factors, and human factors. Also, video quality prediction can be divided into three models. The first model, known as full reference (FR), compares the

∗*Salwa Aqeel Mahdi*

*gosalwyahoo.com@uomustansiriyah.edu.iq*

Communicated by 'sub etitor'

original, distortion-free video with the distorted video for evaluation. The reduced reference (RR) is the second one. In RR, some features about the original videos are available for evaluation. The last one is no reference (NR). In this model, there is no information about original videos, so all features must be taken from the video that has to be qualified.

Deep learning is an advanced machine learning technique that preserves the human mind while analyzing large amounts of data [2]. Many applications use deep learning techniques for different uses. Additionally, a number of deep learning networks, including VGG16 and VGG19, have been developed to extract features from both videos and images [3]. In [4] they Identify and recognise grouped mobile video streams based on a bitstream. It used CNN for the videos to recognise encrypted video streaming from other data streams in mobile applications. This can lead to improved QoE and a better customer experience. However, misclassification occurs due to the application's abilities. In [5], the model provides a higher level of accuracy and requires less computation. A hyper parameter is used in this model to effectively optimize it. Also, the feature extraction is done by using causal convolution. The adaptation model has been tested on two different devices: a mobile phone and a computer. The method in [6] is based on frequency domain depth perception and non-reference assessment. It extracts features in time-frequency domains using DCT. This work [7] uses LSTM ANN to predict the future network throughput. This throughput can be used to estimate the bitrate selection of the video in streaming video for or mobile environments. A transformation method is proposed that transforms the CNN model between domains that is, from spatial local binary mode to extract mathematical statistical features of frequency domain coefficients. Information might vary depending on the mix of features. The goal of selective feature fusion is to choose the image's most pertinent feature [8]. In [9], a model is proposed to evaluate video quality by analyzing the spatial and temporal features and trying to minimize the most of the video information to be used in online video assessment. Researchers conduct an experiment on six databases, employing various deep learning models; including ResNet50.They develop MGQA, an innovative online video assessment approach, by integrating graph networks with MobileNet. A model for predicting video QoE using Multi-Feature Fusion (MFF) was presented. Optimized Learning Models (OLMs) are the optimized neural network models designed for QoE prediction [10]. The inputs to our model are based on the MFF scheme, where the numbers of features that can be combined and provided are between five and six. These features provide a generalized solution to accurately predict perceptual video quality. Shakya et al. [2024] propose a method known as "deep video prior" that quantifies the video by restoring the original from the distorted one [11]. The training phase has to learn the distortion in the video; thus, it is dependent on one reference video to train the model, which limits the generality of the model's ability.

In this paper, we analyze and discover techniques for video quality prediction through deep learning. The next section will give related work for video quality prediction with deep learning. This is followed by a discussion of the results for specific techniques using a new framework.

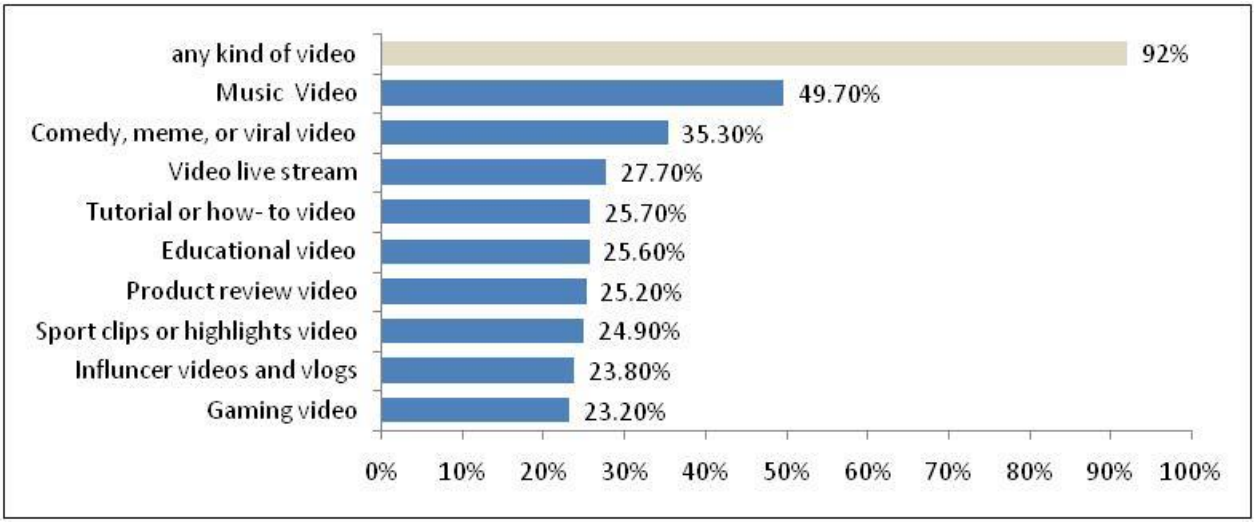


**Figure 1: Videos content types worldwide in 3rd quarter 2023 [1]**

## 2. Related work

Video QoE prediction is a very crucial task for different types of applications, such as video streaming. These applications include stereoscopic video and laparoscopic video, among others. Compression, distortion, or both can lead to the loss of useful information in these videos. This has an influence on user satisfaction or an impact on user happiness, and it can be deadly. All previous surveys and reviews did not attempt to address the deep learning methods used to find the QoE in video multimedia. Telili et al. conducted a review of prediction bitrates in both machine learning and deep learning methods [12]. Also, Zheng et al. provides a survey on the studies that have been done on video quality assessment for user-generated videos, without giving the limitations or weak points of these studies. Moreover, they do not include the evaluation performance [13].

In this paper, we discuss and analyze the most important methods used in the QoE fields for different types of videos. We describe some of the quality prediction techniques in the following subsection. Table 1 give a summary of the techniques, their methods, datasets used, and limitations.

### 2.1 Video Quality Assessor (DeepVQA)

The authors of [14] presented a framework named Deep Video Quality Assessor (DeepVQA). The framework quantifies the video in spatial and temporal visual perception via a convolutional neural network (CNN) and a convolutional neural aggregation network (CNAN).The method is fully referenced, implying that it requires the original video to be without distortion. The framework tries to mimic human perceptions of sensitivity. The evaluation process utilizes two data sets: the LIVE VQA database [15] and the CSIQ VQA dataset [16]. The results for the Live VQA dataset are PLCC 0.8952 and SROCC 0.9152, while for the CSIQ VQA dataset, PLCC 0.9135 and SROCC 0.9123. The model may not be suitable for real-time performance, which is essential for applications such as live video streaming or video conferencing.

### 2.2 Deep Quality of Experience (DeepQoE)

In [17], it introduced a framework for predicting video quality of experience (QoE). The framework has three stages: feature preprocessing, representing learning, and QoE prediction. The generality is achieved through mixed deep learning techniques like word embedding and a 3D conventional neural network (3CD), which are used to extract features in feature preprocessing stages [18]. Different datasets utilize each technique. In the learning stages, a neural network receives these general features as input. In the last stages, a unified prediction for classification and regression takes the output of learning as input. Three datasets are used for evaluation. There are two small datasets, which are WHU-MVQoE2016 [19] and Live-Netflix [20] Video, and a large dataset, Video Set [21]. DeepQoE performs well on regression, even in small datasets. The measurement of accuracy was about 0.88. However, the framework does not compare with other techniques used in deep learning. Also, it does not get higher performance than machine learning.

### 2.3 Cardenas-Angelat model

In [22], it proposed three architectures to predict video quality: single task, multitask, and convolution. These three models are evaluated using simulation experiments in the lab in a mobile environment. It is a fully connected network for single-tasking and multi-tasking. For the convolution, they used CNN. Each model receives four features as input. Initial Loading Delay (ILD) refers to the time in seconds between the initiation of video playback by the user and the actual start of the playback. Secondly, we consider the total time of video stalling (TRB), a result of buffering events. Two statistical measures of the video resolution, its mode (MODEQ) and average (AVGQ), are considered. The CNN model also uses the amount of transmitted data as an input image. The overall performance on

the test set is around 90% accuracy. The limitation of these three models is that there is no generality. The performance of the three models is determined by a specific device environment (on the smart phone) and a specific dataset for evaluations. Moreover, the evaluation results only show each feature separately, not overall features.

### 2.4 Convolutional Neural Networks for Quality of Experience (CNN-QoE)

A new model is proposed to capture the continuous quality prediction of video on different devices. [23] An improvement Temporal Convolutional Network (TCN) [24] is employed to catch the complex dependencies in sequential data. Also, TCN can be implemented in parallel, so it can overcome the computational cost of LSTM. There are many features that can affect video quality in video streaming, such as STRRED, MS-SSIM, PSNR, and STRRED. The experiment is implemented using three datasets. LFOVIA Video QoE [25], LIVE Netflix Video QoE Database and LIVE Mobile Stall Video Database II [26]. The evaluation shows the highest SROCC at 0.885 and the lowest RMSE at 5.27%. Still, TCN has high computation, especially over large datasets.

### 2.5 LSTM-QoE (Long Short-Term Memory-Quality of Experience)

In [27], a method based on a group of LSTM units is proposed to predict adaptive streaming video quality in a continuous environment. The method adopts three features: Short Time Subjective Quality (STSQ), Playback Indicator (PI), and Time Rebuffering (TR). STSQ is a subject perceptual metric that a user is given for a short video segment. There are two values for PI: either the video plays or it undergoes rebuffering. Finally, TR is the duration of time since the last rebuffering happened. The LSTM network is trained using four datasets: the LIVE Netflix Database, the LFOVIA QoE Database the LIVE QoE Database [28], and the LIVE Mobile Video Stall Database-II. Also, mean pooling enhances the performance of LSTM networks. The evaluation showed approximately LCC 0.9. The limitation of this method is the increased computation time as the number of LSTM units grows.

### 2.6 Memory based approach under Three Settling (M-3R)

In [29], a new model is proposed with the name M-3R predictor. The model evaluates QoE in a continuous environment, where each frame is considered. Many features have been extracted: Peak Signal-to-Noise Ratio(PSNR), Structural Similarity Index Measure (SSIM), Multi-scale Structural Similarity (MS-SSIM), Multi-method Assessment Fusion (VMAF), Gradient Magnitude Similarity Deviation (GMSD), STRRED, and Natural Image Evaluation (NIQE). More than one unit of Long-Short Term Memory (LSTM) networks is getting these features as input. A three-setting effect is applied by applying the model to Full Reference (FR), Reduced Reference (RR), and No Reference (NR). The evaluation is done by streaming databases: Live Netflix I, Live NFLX II [30], and Mobile Stall II. The performance result is an approximation of 0.8859 for PLCC, 0.8832 for SROCC, and 0.3453 for RMSE. Since this technique uses a large number of units of LSTM, a high computation complexity is included.

### 2.7 Rapid and Accurate Video Quality Evaluator (RAPIQUE)

The [31] presents a new model to predict the quality of UGC (user-generated content) videos that is based on the no-reference method. The model used two features: natural scene statistics and deep learning features derived from CNN. A vector of both features is built by concatenation. This vector is used as input to get a quality score as output. The model was evaluated using three databases: KoNViD-1k [32], LIVE-VQC [33], and YouTube-UGC [7]. The evaluation results applied to all compound databases. The results show SRCC, PLCC, and RMSE of 0.8070, 0.8229, and 0.3968, respectively. The limitation of the model is the gap differences between RMSE and SRCC, which show more errors than expected for SRCC and PLCC values. Furthermore, it does not compare performance with previous models. In addition, the model has a compression effect.

**Table 1: summary of video QoE prediction using deep learning**

| Name | Author | Features | Method | Datasets | Evaluation measurements | Limitation |
|---|---|---|---|---|---|---|
| **DeepVQA** | Kim 2018 [14] | spatial and temporal features | CNN and CNAN | LIVE VQA CSIQ VQA | Live VQA :PLCC 0.8952 , SROCC 0.9152, CSIQ VQA : PLCC 0.9135 , SROCC 0.9123 | may not suitable for real-time performance since based on full reference |
| **DeepQoE** | Zhang 2018 [17] | Different type | word embeddi ng 3CD | WHU-MVQoE2016 and Live-Netflix Video | accuracy about 0.88 | Not compare with other techniques used in deep learning. |
| **Cárdenas-Angelat model** | Angelat 2019 [22] | ILD, TRB, MODEQ, AVGQ | CNN, FCNN | Database are created | Accuracy approximately 0.09 | No generality ,evaluate only on mobile environment and on one dataset |
| CNN-QoE | Duc 2020 [23] | STRRED,MS-SIM,PSNR, and STRRED. | TCN | LFOVIA LIVE Netflix and LIVE Mobile Stall Video Database II | highest SROCC at 0.885 lowest RMSE at 5.27%. | high computation, especially over large datasets |
| **LSTM-QoE** | Eswara 202 0 [27] | STSQ, PI, TR | LSTM | LIVE Netflix Database, LFOVIA QoE Database, LIVE QoE Database, LIVE Mobile Video Stall Database-II | LCC, SROCC, RMSE<br><br>LFOVIA 0.858, 0.808, 0.0864<br><br>Netflix Video 0.802, 0.714, 0.0778 Respectively | increased computation time as the number of LSTM units grows |
| **M -3R** | Ghosh 2021 [29] | PSNR,SSIM,MS-SSIM, VMAF, GMSD , STRRED, ,NIQE. | LSTM | Live Netflix I, Live NFLX II, and Mobile Stall II | 0.8859 for PLCC, 0.8832 for SROCC, and 0.3453 for RMSE | high computation |
| **Name** | **Author** | **Features** | **Method** | **Datasets** | **Evaluation measurements** | **Limitation** |
| **RAPIQUE** | Tue 2021[31] | natural scene statistics and deep learning features | CNN | KoNViD-1k , LIVE-VQC , YouTube-UGC | SRCC 0.8070, PLCC 0.8229, RMSE , 0.3968, | compare performance with previous model |
| **No-Reference Quality Assessment** | Gu 2022 [34] | frames in the DCT frequency | CNN | NAMA3DS1-COSPAD1 , WaterlooIVC 3D Video Phase I , QI- SVQA database | PLCC 0.9520 , SROCC 0.9458 , RMSE 0.2994 | performance-focused frequency domain |
| **DeSVQ** | Ghosh 2022 [38] | Frames, PSNR MSSSIM, STRRED, VMAF | LSTM, CNN | LIVE Netflix I, LIVE NFLX II, Mobile Stall II DeSVQ the Mobile Stall II dataset. | LCC (0.8988), SROCC (0.8936), and least RMSE (0.363) on Mobile Stall II | high computation |
| **MO-QoE** | Ghosh 2022 [10] | MFF | ANN , FNN | LIVE, VQEG HD3, VQEG, HD4 , LIVE Netflix , LIVE NFLX II , Waterloo , CSIQ , LFOVIA | highest ROCC is 0.9328, the PLCC highest is 0.9360 | No compare with other models |
| **3S-3DCNN** | Islam 2022 [39] | spatial, motion, and depth features | 3D CNN | LFOVIAS3DPh2 , NAMA3DS1-COSPAD1 | RSME 0.2757 for NAMA3DS1-COSPAD1 SROCC to exceed 98% for in LFOVIAS3DPh2. | the evaluation is restricted to only two datasets |

| | | | | | | |
|---|---|---|---|---|---|---|
| **VQP-Net** | Khan 2022 | frame and temporal vector of quality | FCCN<br><br>ResNet | LVQ database | PLCC and SROCC of 0.9899 and 0.9388,. | limited scope only one database. Not considered not the compression artifact |
| **Deep ISVM** | Elwerghemmi 2023 | visual patterns, motion, and temporal | CNN | Poqemon Database, LIVE-Netflix Video Database, and LFOVIA Database | accuracy for Poqemon about 86.35, LIVE-Netflix Video is 58.33, and LFOVIA is 75.29 | the variances accuracy result between datasets |
| **CLN-RLN** | Liu, 2023 | KPI | FCN | Database are created | without the split RMSE about 0.09757 with splitting RMSE 0.0175150 | no generalization. unrealistic linearity between QoS and QoE. |

### 2.8 No-Reference Quality Assessment Based on Deep Frequency Perception

In (Gu et al. 2022), they proposed a method to quantify stereo video using frequency domain characterization, with no reference video needed [34]. The video resolution is reduced to solve the problem of the large communication bandwidth, which will lead to the loss of some information in the spatial domain. To overcome this loss that affects accuracy, features are extracted from video frames in the frequency domain using DCT. The CNN model is used to extract the mathematical statistics of frequency coefficients, which reduces the size of the data since high-frequency information has a high impact value. The experiment uses three datasets: the NAMA3DS1-COSPAD1 [35] database, the Waterloo IVC 3D [36] Video Phase I database, and the QI-SVQA database [37]. The evaluation of this model was PLCC, SROCC, and RMSE, which were about 0.9520, 0.9458, and 0.2994, respectively. The limitation of this model is the performance-focused frequency domain features of stereo video.

### 2.9 Deep Learning-Based Streaming Video QoE Estimation (DeSVQ)

A deep learning method is described in [38]. It uses a framework that combines Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) networks. They combine different features that were extracted in two steps and show how the QoE is affected by complex dependencies during the prediction process. In the initial stages, CNN directly extracts the features from the frames in the distorted videos. The LSTM network then sequentially maps them to QoE scores. In another stage, the LSTM network explores the temporal dependencies using the objective (numerical) features. Peak-Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index (MSSSIM), Spatio-Temporal Reduced Reference Entropic Differencing (STRRED), and Video Multi-method Assessment Fusion (VMAF) are five of these features. The output from both stages is linearly combined and fed to the decision trees. The DeSVQ architecture is validated on three different datasets. LIVE Netflix I, LIVE NFLX II, and Mobile Stall II DeSVQ get the highest LCC (0.8988), SROCC (0.8936), and least RMSE (0.363) on the Mobile Stall II dataset. One limitation of this method is its high computation since LSTM is considered sequential processing.

### 2.10 MO-QoE model

In [10], a model for predicting video QoE using Multi-Feature Fusion (MFF) was presented. Optimized Learning Models (OLMs) are the optimized neural network models designed for QoE prediction. The OLM algorithm is implemented using neural networks (ANN and FNN). These two neural networks are considered the foundation for deep learning. An Adaptive Moment estimation and Batch Gradient Descent algorithm are used to update the weights and biases of the learning models to their optimal values. The inputs to our model are based on the MFF scheme, where the numbers of features that can be combined and provided are between five and six. These features are given a generalized solution to predict perceptual video quality effectively. Many databases have been used for evaluating our proposed model, including LIVE, VQEG HD3, VQEG HD4, LIVE Netflix, LIVE NFLX II, Waterloo, CSIQ and LFOVIA. The accuracy is computed using SROCC; the highest one is 0.9328, and the PLCC is 0.9360.

### 2.11 Three Stream the Three-Dimension Convolution Neural Network (3S-3DCNN)

In [39] a non-reference technique for assessing stereoscopic video is proposed. Two cameras position themselves at an distance to create stereoscopic video [40]. The 3D CNN architecture is based on spatial and temporal features between sequence frames, disparity, and motion information. To assess the quality of stereo video, a 3-Stream 3D CNN (3S-3DCNN) extracts spatial, motion, and depth features. These extracted features are then fed into fully connected layers for regression. Two datasets are used for evaluation: LFOVIAS3DPh2 [41] and NAMA3DS1-COSPAD1. Also, a modified 3S-3DCNN model called 3S-3DCNN-clean is proposed and applied. The experimental evaluation showed RMSE to be 0.2757 for dataset NAMA3DS1-COSPAD1 and SROCC to exceed 98% for distortion video in LFOVIAS3DPh2. The evaluation limits the 3S-3DCNN model to only two datasets, despite its impressive performance.

### 2.13 Video Quality Prediction Network (VQP-Net)

In [42], a new framework is proposed that predicts the quality of laparoscopic videos. Laparoscopic videos are videos that are recorded during surgical operation procedures and are used for training students. VQP-Net applies the previous learning models sequentially to each frame. It will give a temporal vector of quality score that will be applied to FCCN to get the predicted result for video. There are two learning approaches: transfer learning and end-to-end learning. The transfer learning used in the pre-trained FQP-ResNet model will only be applied to the FCNN. This will reduce the complexity of training computation time. The end-to-end learning approach applies the weight update to the entire model, so the same loss function is used to make it more homogenous. The experiment was applied to the LVQ database [43], which contains videos of cholecystectomy surgeries with distortion. The results show PLCC and SROCC values of 0.9899 and 0.9388, respectively. The limited scope of the work is that it applied only to one database. In addition, it is based on the distortion caused by different noises but not the compression artifact.

### 2.12 Deep Incremental Support Vector Machine (Deep ISVM)

Elwerghemmi et al. proposed a method that combines both deep learning networks and machine learning algorithms to predict video quality [44,45]. The aim is to continuously quantify the perceived QoE of streaming video. The framework consists of two levels. The first level is based on the DeepQoE model [17] for reprocessing and feature extraction from videos. The deep learning network extracts discriminative features for video using CNN architecture. These features are selected from video frames or sequences, visual patterns, motion, and temporal. The second level, based on the stacking model, is the ensemble method, where many classification models are combined via a meta-classifier [46]. The incremental support vector machine (ISVM) is a multiclass model that is based on the stacking method to enhance prediction. A five-level ISVM classifier gets output from the first level, and the highest output will fire the prediction before the threshold. This method, along with ISVM, also allows the model to adapt to new data without retraining the entire network. The ISVM incrementally updates its decision boundaries as new samples arrive. Three databases are applied: the Pokémon Database [47], the LIVE-Netflix Video Database, and the LFOVIA Database. The overall performance of the method shows that its maximum accuracy for the Pokémon data set is about 86.35, for the LIVE-Netflix Video dataset it is 58.33, and for LFOVIA it is 75.29. The model was evaluated only on three databases, with some having very low accuracy. A refinement to the deep ISVM model needs to decrease the variance accuracy result.

### 2.14 Classification Learning Network-Regression Learning Network (CLN-RLN)

Liu et al. [48] proposed a model that is based on the QoE/QoS mapping. QoE predicts from QoS collected every millisecond. A number of network performance measurements and key performance indicators (KPIs) are used to show the connection between QoE and QoS in wireless networks. The model is built with two deep-learning neural

networks. The first-level FCN, known as the classification learning network (CLN), accurately classifies the input data into various categories. Then, the classification results will be processed by the second-level FCN, the regression learning network (RLN), which performs regression to predict the quality. This clustering will help to decrease the variance between data in one batch and increase performance. Also, the datasets are provided by the MATLAB simulator. The proposed model has a root mean square error (RMSE) of about 0.09757 without the split data technique. However, the model with splitting gets 0.0175150. The proposed method applied and examined only a specific dataset, which restricted generalization. Also, it is based on the PKI metric, which, although unrealistic and impractical, assumes linearity between QoS and QoE.

## 3. Summary and Analysis for related work

The previous sections presented various deep learning methods for predicting video quality. These deep learning methods vary in difficulty, such as convolutional neural networks (CNNs), long short-term memory (LSTM) models, and more complex designs like 3D CNNs, hybrid fusion models, and ensemble learning. DeepVQA and CNN-QoE used CNNs and special methods to understand both space and time in a fully referenced way. LSTM-QoE, DeSVQ, and M-3R, on the other hand, concentrated on capturing temporal dependencies using LSTM networks, especially in the context of continuous streaming scenarios. The models that use LSTM often come with increased computational costs due to their sequential processing nature.

The newest models, such as DeepQoE, MO-QoE, and DeepISVM, emphasize feature fusion and optimization by integrating different perceptual and objective quality metrics to improve generalization across datasets. Models like RAPIQUE and 3S-3DCNN present no-reference strategies, making them suitable for real-world applications where undistorted reference videos are unavailable to compare. Certain models, such as VQP-Net, concentrate on identifying the quality of laparoscopic videos. The CLN-RLN model aimed to predict the quality of experience in scenarios with limited network resources by utilizing quality of service parameters. Despite achieving high performance in correlation metrics (PLCC, SROCC), many models still face limitations due to their dataset dependency, computational complexity, or application scope. In general, even though deep learning has greatly improved how we understand video quality, future efforts should focus on fixing issues with how well these models work in different situations, their ability to work in real time, and their strength across various fields.

## 4. Discussion

Three models are selected to be analyzed. We examine three models: CNN-QoE, LSTM-QoE, and DeSVQ. The three models are applied to two different data sets, which are the Live Netflix Database, and Live Mobile Video Stall Database-II. LIVE Netflix Database has 120 videos, LFOVIA has 36 videos, and LIVE Mobile Video Stall Database II has 174 videos. While DeSVQ is applied Live Netflix Database, Database, and Live Mobile Video Stall Database. Three measurements are used: LCC (Linear Correlation Coefficient), SROCC (Spearman Rank Order Correlation Coefficient), and RMSE (Root Mean Square Error). The LCC and SROCC have values between 1 and -1. LCC measures the correlation between predicted and actual predictions, like mathematical correlation [49]. Also, SROCC measures the correlation between predicted and actual predictions in terms of human perception. The two following equation is given the :

$$\mathrm{SROCC} = 1 - \frac{6\ \sum_{i=1}^{n}\ (y_i - \hat{y}_i)^2}{n(n^2 - 1)}$$

$$\mathrm{LCC} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n number of data value; is $y_i$ denotes the actual value, $\bar{\hat{y}}_i$ represents the predicted value. $\bar{\hat{y}}$ is the mean of the predicted values, and $\bar{y}$ is the mean for the actual values Where n is the amount of data set

On the other hand, RMSE measures the differences between predicted and actual values. As the value of RMSE decreased, the prediction was better [50].

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

Where n number of data value; is $y_i$ denotes the actual value, $\hat{y}_i$ represents the predicted value.

**Table 2 Performance result of three models LSTM-QoE, CNN-QoE, and DeSVQ**

| Dataset | Measurement | **Models** | | |
|---|---|---|---|---|
| | | LSTM-QoE | CNN-QoE | DeSVQ |
| Mobile Video Stall Database-II | LCC | 0.939 | 0.892 | 0.8988 |
| | SROCC | 0.936 | 0.885 | 0.8936 |
| | RMSE | 0.05702 | 0.0536 | 0.352 |
| Live Netflix I | LCC | 0.802 | 0.873 | 0.8935 |
| | SROCC | 0.714 | 0.878 | 0.8908 |
| | RMSE | 0.0778 | 0.0527 | 0.327 |

Table 2 for the data set Mobile Video Stall Database-II shows that LSRM-QoE performs better than both CNN-QoE and DeSVQ. This is due to the fact that it is better at modeling temporal dependencies in stalling that is common in mobile video. For the RMSE metric, LSTM-QoE and CNN-QoE are almost the best at getting the minimum value. In addition, CNN-QoE has lower LCC/SROCC than LSTM-QoE. DeSVQ has a higher RMSE. but it has better correlation between predicted value and MOS when compared with CNN-QoE. For dataset Live Netflix I DeSVQ is performed in LCC and SROCC, which indicates that it is more generalized on this type of data. But still has a higher RMSE when compared with two models. CNN-QoE achieves a lower RMSE, which indicates a good prediction but has a lower correlation. LSTM-QoE performs poorly in this dataset.

The comparison shown in figure 2 shows how three QoE prediction models—LSTM-QoE, CNN-QoE, and DeSVQ—perform on two standard datasets: Mobile Video Stall Database-II and Live Netflix I. Mobile Video Stall Database-II and Live Netflix I. LSTM-QoE shows better performance in both LCC and SROCC on the Mobile Video Stall Database II, which mostly demonstrates stall-related conditions, therefore suggesting its ability to capture temporal quality variations. The ability of LSTM to model sequential dependencies is a key feature. Thus, LSTM provides this benefit, which makes it especially useful for material rich in stalls. CNN-QoE somewhat surpasses LSTM-QoE in terms of RMSE. That CNN-QoE suggests more accuracy in numerical score predictions, but the difference is minor. Although competitive in correlation measures, DeSVQ has a much greater RMSE, suggesting less consistent raw score predictions. DeSVQ outperforms the others in both LCC and SROCC in the Live Netflix dataset, which consists of an expanded range of quality degradations, including bitrate variations and compression artifacts. It means that DeSVQ is more suitable for capturing more complex quality variances outside of temporal distortions. CNN-QoE shows consistency in correct score estimation by maintaining its trend of obtaining the lowest RMSE across datasets. On this dataset, LSTM-QoE does not perform well due to its poor ability to handle non-temporal deviations. The study demonstrates that, whereas LSTM-QoE appears in situations with temporal quality falls, CNN-QoE and DeSVQ provide greater generalisations and accuracy in more varied video streaming environments.

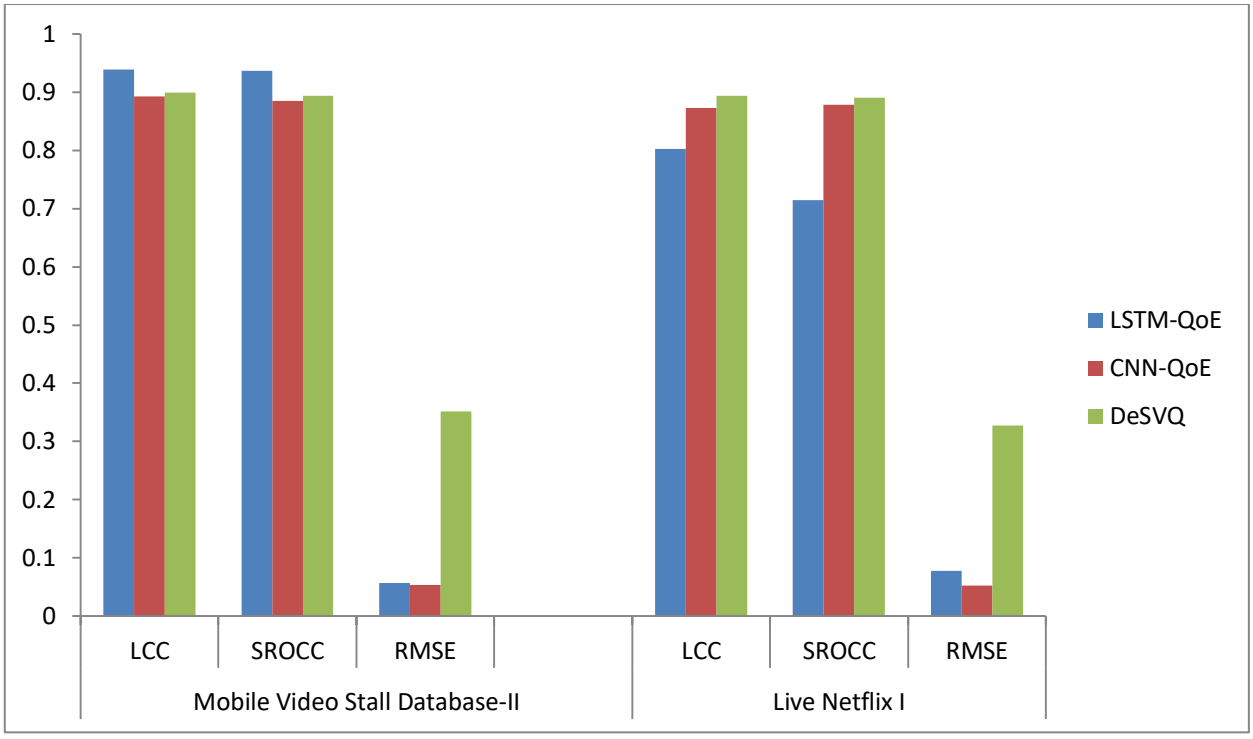


**Figure 2: Evaluation of CNN-QoE, LSTM-QoE and CLN-RLN**

All the previous work is depending on subjective metrics, which are MOS in training the deep learning model. A new framework is proposed that integrates the subjective metric (MOS) with an objective metric that can be computed from the video itself. The newly created metric can be applied as new values to start the training. Each metric will be tuned with weight since each feature has a different effect on different video datasets. Additionally, we must adjust the updated weight to address the issue of forgetting.

Additionally, this new metric divides the data set into multiple groups. This new method is used for the idea of augmented labelling to train the new models. The augmented labelling is used to increase the accuracy of the model and help overcome errors that sometimes appear due to human views. Each feature will be tuned with weight since each feature has a different effect on different video datasets. Moreover, we must adjust the updated weight to address the issue of forgetting.

## 4. Conclusion

This paper investigates quality prediction methods. We have clarified many models, their methods, datasets, and limitations. We have also explained the performance of three models: LSTM-QoE, CNN-QoE, and DeSVQ and the reasons behind their outcomes. LSTM-QoE is better in temporary-dependence features due to its sequence-modeling strength. CNN-QoE regularly provides low RMSE, making it more suitable for precise MOS prediction. DeSVQ offers strong performance on more complex or less temporally obvious distortions. In addition, a new framework is proposed. This new framework is based on the limitations of the three models. Moreover, it attempts to make more general models that can predict different types of video from different datasets.

## References

1. Digital 2024, "Global Overview Report," STATISTA, p. 109, (2024). [Online]. Available: https://www.statista.com/statistics/1254810/top-video-content-type-by-global-reach
2. M. Haqi Al-Tai, B. M. Nema, and A. Al-Sherbaz, "Deep Learning for Fake News Detection: Literature Review," AL-MUSTANSIRIYAH JOURNAL OF SCIENCE, vol. 34, no. 2, pp. 70–81, (2023), doi: 10.23851/mjs.v34i2.1292.

3. H. S. Ibrahim, N. M. Shati, and A. A. Alsewari, "A Transfer Learning Approach for Arabic Image Captions," AL-MUSTANSIRIYAH JOURNAL OF SCIENCE, vol. 35, no. 3, pp. 81–90, (2024) , doi: 10.23851/mjs.v35i3.1485.
4. D. Chmieliauskas and Š. Paulikas, "Video Stream Recognition Using Bitstream Shape for Mobile Network QoE," Sensors, vol. 23, p. 2548, (2023), doi: 10.3390/s23052548.
5. B. Mahaboob and S. A. Kalaiselvan, "Experimental Investigation Based on Services of Video Streaming using Deep Neural Network for Continuous QoE Prediction," Journal of Theoretical and Applied Information Technology, vol. 101, no. 5, pp. 1954–1961, Mar. (2023). [Online]. Available: www.jatit.org.
6. F. Gu and Z. Zhang, "No-Reference Quality Assessment of Stereoscopic Video Based on Temporal Adaptive Model for Improved Visual Communication," Sensors, vol. 22, no. 21, p. 8084, (2022).
7. Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC Dataset for Video Compression Research," in PROC. IEEE INT. WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING (MMSP), Kuala Lumpur, Malaysia, (2019) , pp. 1–5, doi: 10.1109/MMSP.2019.8901772.
8. N. M. Khassaf and S. H. Shaker, "Image Retrieval based Convolutional Neural Network," AL-MUSTANSIRIYAH JOURNAL OF SCIENCE, vol. 31, no. 4, pp. 43–54, (2020), doi: 10.23851/mjs.v31i4.897.
9. J. Yan, L. Wu, Y. Fang, X. Liu, X. Xia, and W. Liu, "Video Quality Assessment for Online Processing: From Spatial to Temporal Sampling," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 12_Part_2, pp. 13441–13451, Dec. (2024), doi: 10.1109/TCSVT.2024.3450085.
10. M. Ghosh and C. Singhal, "MO-QoE: Video QoE using multi-feature fusion based Optimized Learning Models," Signal Processing: Image Communication, vol. 107, (2022) , p. 116766. ISSN: 0923-5965
11. S. N. Shakya and P. Kancharla, "Deep priors for video quality prediction," arXiv preprint*, arXiv:2410.22566, 2024. doi: 10.48550/arXiv.2410.22566.
12. A. Telili, A. Ksentini, Y. Hadjadj-Aoul, V. Pégon, and A. C. Begen, "Bitrate ladder prediction methods for adaptive video streaming: A review and benchmark," IEEE Transactions on Multimedia, vol. 24, no. 8, pp. 2225–2236, (2022) , doi: 10.1109/TMM.2022.3158572.
13. Q. Zheng, Y. Wang, J. Zhang, K. Zhang, and W. Lin, "Video Quality Assessment: A Comprehensive Survey," arXiv preprint arXiv:2412.04508, Dec. (2024).
14. W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in European Conference on Computer Vision (ECCV), Springer, Cham, (2018), pp. 219-234. doi: 10.1007/978-3-030-01237-3_17.
15. K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427–1441, (2010).
16. Laboratory of Computational Perception & Image Quality, Oklahoma State University, "CSIQ video database," (2013).
17. H. Zhang, H. Hu, G. Gao, Y. Wen, and K. Guan, "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction," IEEE Trans. Multimedia, vol. 22, no. 12, pp. 3210–3223,( 2020) , doi: 10.1109/TMM.2020.2973828.
18. H. Zhang, H. Hu, G. Gao, Y. Wen, and K. Guan, "DeepQoE: A Unified Framework for Learning to Predict Video QoE," in 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, (2018), pp. 1-6.
19. Y. Zhang, M. Yuan, and Z. Chen, "WHU-MVQoE2016: A quality of experience dataset for mobile video research," WHU Tech. Rep., Dec. (2016).
20. C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron and A. C. Bovik, "LIVE Netflix Video Quality of Experience Database," Online: http://live.ece.utexas.edu/research/LIVE_NFLXStudy/index.html, (2016).
21. H. Wanget al., "VideoSet: A large-scale compressed video quality dataset based on jnd measurement,"J. Visual Commun. Image Representation, vol. 46, pp. 292–302, (2017).
22. C. Cárdenas-Angelat, J. B. Polglase, C. J. Vaca-Rubio, and M. C. Aguayo-Torres, "Application of Deep Learning Techniques to Video QoE Prediction in Smartphones," in 2019 European Conference on Networks and Communications (EuCNC), Valencia, Spain, (2019), pp. 252-256. doi: 10.1109/EuCNC.2019.8801974.
23. T. N. Duc, C. M. Tran, P. X. Tan, and E. Kamioka, "Convolution Neural Networks for Continuous QoE Prediction in Video Streaming Services," IEEE Access, vol. 8, pp. 116268-116278, (2020). doi: 10.1109/ACCESS.2020.3004125.
24. S. Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint, arXiv:1803.01271, (2018). [Online]. Available: http://arxiv.org/abs/1803.01271.
25. N. Eswara, K. Manasa, A. Kommineni, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya, "A

Continuous QoE Evaluation Framework for Video Streaming Over HTTP," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3236–3250, nov (2018). [Online]. Available: https://ieeexplore.ieee.org/document/8013810/

26. D. Ghadiyaram, J. Pan, and A. C. Bovik, "A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 183–197, (2019). [Online]. Available: http://ieeexplore.ieee.org/document/8093636/
27. N. Eswara et al., "Streaming Video QoE Modeling and Prediction: A Long Short-Term Memory Approach," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 3, pp. 661-673, Mar. (2020). doi: 10.1109/TCSVT.2019.2895223
28. C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, "Modeling the time-varying subjective quality of HTTP video streams with rate adaptations," IEEE Transactions on Image Processing, vol. 23, no. 5, pp. 2206–2221, May (2014).
29. M. Ghosh and C. Singhal, "M-3R: A Memory Based Approach for Streaming QoE Prediction under 3R settings," in 2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Hyderabad, India, (2021), pp. 432-437. doi: 10.1109/ANTS52808.2021.9936944.
30. LIVE -Labaratory for Image and Video Quality Engineering, an Image Quality Assessment Database. Available: http://live.ece.utexas.edu/research/LIVE NFLX II/live nflx plus.html.
31. Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," IEEE Open Journal of Signal Processing, vol. 2, pp. 425–440, (2021).
32. V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szir´anyi, S. Li,and, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX), (2017), pp. 1–6.
33. Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," IEEE Trans. Image Process, vol. 28, no. 2, pp. 612–627, (2018).
34. F. Gu, & Z. Zhang, "No-Reference Quality Assessment of Stereoscopic Video Based on Temporal Adaptive Model for Improved Visual Communication," Sensors, (2022), 22(21), 8084. https://doi.org/10.3390/s22218084.

35. M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P.L. Callet, J. Gutierrez, and N. Garcia, “NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences.” In Proceedings of the 2012 Fourth International Workshop on Quality of Multimedia Experience, Melbourne, Australia,5–7 July (2012).
36. J. Wang, S. Wang, Z. Wang, “Quality prediction of asymmetrically compressed stereoscopic videos.” In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September (2015).
37. F. Qi, T. Jiang, X. Fan, S. Ma, and D. Zhao, “Stereoscopic video quality assessment based on stereo just-noticeable difference model,” In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September (2013).
38. M. Ghosh, R. Wayal, and C. Singhal, "DeSVQ: Deep Learning Based Streaming Video QoE Estimation," in Proceedings of the 23rd International Conference on Distributed Computing and Networking (ICDCN 2022), January 4–7,)( 2022), Delhi, India, ACM, New York, NY, USA, 7 pages. doi: 10.1145/3491003.3491023.
39. M. B. Islam, "Three-Stream 3D deep CNN for no-Reference stereoscopic video quality assessment," Intelligent Systems with Applications, Jan. (2022). doi: 10.1016/j.iswa.2021.200059.
40. D. R. Bull and F. Zhang, Chapter 2 "The human visual system," in Intelligent Image and Video Compression (Second Edition), D. R. Bull and F. Zhang, Eds. Academic Press, (2021), pp. 17-58. ISBN: 9780128203538. doi: 10.1016/B978-0-12-820353-8.00011-6.
41. B. Appina, S. V. R. Dendi, K. Manasa, S. S. Channappayya, and A. C. Bovik, "Study of subjective quality and objective blind quality prediction of stereoscopic videos," IEEE Transactions on Image Processing, vol. 28, no. 10, pp. 5027–5040, Oct.( 2019).
42. Z. A. Khan, A. Beghdadi, M. Kaaniche, F. Alaya-Cheikh, and O. Gharbi, "A neural network based framework for effective laparoscopic video quality assessment," Computerized Medical Imaging and Graphics, vol. 101, p. 102121, Oct. (2022). doi: 10.1016/j.compmedimag.
43. Z. A. Khan, A. Beghdadi, F. A. Cheikh, M. Kaaniche, E. Pelanis, R. Palomar, A. A. Fretland, B. Edwin, O. J. Elle, Towards a video quality assessment based framework for enhancement of laparoscopic videos, in: SPIE 33Medical Imaging: Image Perception, Observer Performance, and Technology Assessment, Vol. 11316, (2020), p. 113160P.
44. R. Elwerghemmi, M. Heni, R. Ksantini, and R. Bouallegue, "Online QoE Assessment Model Based on Incremental Stacked Multiclass Classifier," International Journal of Computing and Digital Systems, (2023).
45. R. Elwerghemmi, M. Heni, R. Ksantini, and R. Bouallegue, "An Efficient Stacked Deep Incremental Model for Online Streaming Video QoE Prediction,"International Journal of Computing and Digital Systems, vol. 13, no. 1, pp. 1-12, (2023). doi: 10.12785/ijcds.1301119.
46. R. Elwerghemmi, M. Heni, R. Ksantini, and R. Bouallegue, "Online QoE Prediction Model Based on Stacked Multiclass Incremental Support Vector Machine," in 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, (2019), pp. 1-5. doi: 10.1109/ICMSAO.2019.8880302.
47. Poqemon-QoE-Dataset," Github. [Online]. Available: https://github.com/Lamyne/Poqemon-QoE-Dataset
48. C. Liu, X. Chen, X. Wang, X. Xie, and Z. Guo, "QoE Assessment Model Based on Continuous Deep Learning for Video in Wireless Networks," IEEE Transactions on Mobile Computing, vol. 22, no. 6, pp. 3619-3633, (2023).
49. A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," arXiv preprint arXiv:1809.03006, Sept. (2018).
50. V. Plevris, G. Solorzano, N.P. Bakas, and M.E.A Seghier, “Investigation of performance metrics in regression analysis and machine learning-based prediction models, ” In ECCOMAS Congress (2022).