Leveraging Multi-View Learning for Quality of Experience Prediction Models

Matteo Fratta[1], Simone Porcu[1,2], Giulia Martinelli[3], Alessandro Floris[1,2], and Luigi Atzori[1,2]
[1]DIEE, University of Cagliari, 09123 Cagliari, Italy
[2]CNIT, University of Cagliari, 09123 Cagliari, Italy
[3]DISI, University of Trento, 38123 Trento, Italy
{matteo.fratta, simone.porcu}@unica.it, giulia.martinelli-2@unitn.it, {alessandro.floris84, l.atzori}@unica.it

***Abstract*—Accurate models are necessary for the continuous estimation of the Quality of Experience (QoE), which is crucial for delivering successful multimedia services to end-users. These models are developed from subjective test data, which very often provide only specific aspects of the experience, i.e., a *partial view* (PV). Each PV conveys a relationship between a specific set of influence factors and the perceived QoE, limiting the applicability of the derived model to other application scenarios not considered in the initial subjective tests. To extend the applicability of the developed models, this paper introduces a multi-view (MV) learning framework that enhances QoE prediction by integrating complementary information from multiple perspectives obtained from different subjective tests. We leverage a fully connected deep neural network with two initially independent branches and an intermediate fusion layer to combine insights from separate feature sets, improving predictive accuracy while preserving data privacy. Our model is trained on a synthetic data set derived from the TID2008 image database, ensuring a controlled yet representative evaluation environment. On the one hand, the results demonstrate that the MV technique outperforms all PV configurations. On the other hand, the MV approach achieves QoE estimation performance comparable to the single-view (SV) model, in which one single branch analyzes the full set of impact factors. In particular, the largest performance gain (6.15% – 142.69%) across most evaluation metrics occurred when the input data set is equally divided between the two separate views.**

***Index Terms*—Quality of Experience, Multi-view Learning, Neural Network, QoE prediction.**

I. Introduction

Traditional Quality of Experience (QoE) assessment methods rely heavily on subjective tests involving human participants who are asked to rate their perceived quality after exposure to different contents or stimuli. The collected individual scores are often aggregated to compute the Mean Opinion Score (MOS), which can be used to train objective QoE estimation methods that leverage mathematical algorithms to predict the QoE without direct human intervention [1].

Subjective tests retain inherent advantages over objective QoE estimation. In particular, they provide invaluable insight into the nuances of human perception that might be potentially overlooked by objective measures, such as the impact of complex interactions between various quality factors and the influence of individual differences. Moreover, subjective scores are particularly valuable in identifying relevant quality attributes and dimensions that may not be readily apparent through purely quantitative analysis. Despite these aspects, the main limitation of subjective measures remains the inability to be used in real-time QoE-aware service management, which is one of the main reasons for assessing the user's QoE [2].

To build accurate objective QoE models, a substantial amount of (extensive) subjective test data is required. Usually, the collected data refers to a specific application scenario to which the developed model is only applicable. Additionally, each subjective study considers a limited number of QoE influence factors (IFs), making it hard to define a comprehensive QoE prediction model. Accordingly, not only is the derived model applicable to the specific application scenario (e.g., video streaming, audiovisual call), but it can only be fed with the same IFs preliminarily included in the subjective tests.

Multi-view (MV) representation learning offers a potential solution to this issue by exploiting data from different views (i.e., diverse feature sets or data sources usually containing complementary information, often referred to as *Partial Views*) to learn more comprehensive representations than those of single-view (SV) traditional machine learning (ML) algorithms (e.g., kernel machines, support vector machines) [3]. MV learning enhances SV learning by assigning a unique function to each view and optimizing all functions together to leverage the redundant views of the same input data, thereby enhancing learning accuracy, robustness, and generalization capabilities [4]. While MV learning has become a very promising topic with wide applicability, such as multimedia analysis [5], [6] and human activity recognition [7], [8], its application in QoE modelling is still limited in the literature [9].

This paper proposes an MV-based collaborative approach to extend the applicability of QoE models by integrating complementary information from multiple views. We leverage a fully connected deep neural network with two initially independent branches and an intermediate fusion layer to combine insights from separate feature sets, considering diverse image distortions, intending to improve predictive accuracy while preserving data privacy. Without losing the general applicabil-

This work has been partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP C29J24000300004, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") and by the European Union under the NRRP of NextGenerationEU, "Sustainable Mobility Center" (Centro Nazionale per la Mobilità Sostenibile, CNMS, CN_00000023).

ity of the proposed approach to other application settings, our model has been specifically trained to estimate the quality of images on a synthetic dataset derived from the TID2008 image database [10], ensuring a controlled yet representative evaluation environment. On the one hand, the results demonstrate that the MV approach achieves QoE estimation performance comparable to, and even superior to, a neural network trained on the full set of features (single-view). On the other hand, the MV technique significantly outperforms all partial views configurations. Finally, we systematically analyze the impact of different feature partitions, highlighting the configurations that yield optimal performance. It is worth emphasizing that, given the synthetic nature of the training dataset, the primary objective of our work is to demonstrate the effectiveness of the MV learning paradigm in the context of QoE estimation, rather than to propose a definitive predictive model.

The paper is structured as follows. Section II discusses related work. In Section III, we present our MV approach for QoE modelling, whereas Section IV outlines the implementation and training processes. Section V assesses the model performance, while in Section VI we draw our conclusions.

II. Related Work

MV learning has proven effective across various domains. In [11], one of the earliest schemes for MV learning, the co-training algorithm, was proposed, which trains alternately to maximize the mutual agreement on two distinct views of the unlabeled data. It was proposed to improve the accuracy of the Web document classification task. Besides applications involving text or natural language processing (e.g., e-mail classification [12]), MV has also found applications in the field of computer vision and multimedia analysis. A Siamese convolutional neural network (CNN) is proposed in [13] to force the MV expression recognition model to learn the same features as the frontal expression recognition model. To further enhance recognition accuracy, a multi-task learning framework concerning head pose estimation was also adopted. A deep neural network-based MV method is proposed in [5] for facial emotion recognition, which can learn a set of optimal features for classifying the facial expressions across different facial views, resulting in improved performance compared to existing methods on two non-frontal facial expression databases. In [6], MV feature learning is employed to jointly exploit tampering boundary artifacts and the noise view of the input image for detecting image manipulation. The authors of [7] formulated a multi-view dynamic image fusion task as a multi-instance learning problem, i.e., they considered the dynamic images derived from different virtual imaging viewpoints as the viewpoint instances for 3-D action characterization. This approach demonstrated the superiority and generality in cross-view 3-D action recognition tasks, outperforming literature methods.

Despite the studies mentioned above demonstrating that MV learning can lead to enhanced performance for prediction models in diverse domains, limited studies have employed this learning approach for QoE modelling. Some solutions focused on the fusion of data representations originating from different datasets have been investigated. Distributed ensemble learning and vertical federated learning techniques were employed in [14] to combine predictions of models independently trained on some or all diverse feature sets, respectively. However, the experiments that were conducted only considered binary classifications by quantizing MOS scores into two classes. A split learning approach is proposed in [15], aimed at training a QoE estimation model based on decentralized observations, e.g., in the Application server and Network node, which can still be collected and trained on local neural network models. Nonetheless, the lack of datasets made it unclear the potential of this approach. To the best of the authors' knowledge, the study in [9] is the only one utilizing MV learning for QoE modeling. A dataset concerning the QoE of Web browsing sessions was artificially partitioned into two distinct views and used to train initially separated neural networks, one on each view. An intermediate fusion layer was then used to allow both networks to exchange pre-processed information, implementing the MV learning. The results revealed that the MV approach yielded QoE estimation performance comparable to the SV model and significantly outperformed the performance of the two separated networks.

III. Multi-view approach for QoE modeling

The literature analysis in Section II has outlined that further research is needed to determine if MV learning is feasible and effective for QoE prediction and to explore its potential benefits compared to other learning approaches. The necessity of adopting such an approach arises from the inherently multifaceted nature of QoE assessment, which is influenced by a broad spectrum of IFs. Indeed, existing targeted studies often focus on specific subsets of IFs, leading to models that fail to account for the broader picture that shapes the overall user experience. As a result, their predictive accuracy is suboptimal, as they consider only a partial representation of the problem and overlook critical cross-domain effects. Furthermore, even when different studies consider IFs of a similar nature, inconsistencies in measurement methodologies or variation ranges can hinder the direct comparison and integration of their findings.

The MV learning paradigm presents an encouraging framework to address these challenges, as it enables the fusion of complementary insights from independent models, often referred to as *Partial Views* (PVs), into a unified model that enhances predictive accuracy and generalizability. A practical example, which is also considered in [15], is the QoE dependence on both application-layer IFs (e.g., video resolution, frame rate) and network-layer IFs (e.g., latency, packet loss). Independent QoE estimations conducted separately by over-the-top (OTT) application providers or Internet Service Providers (ISPs) can be incomplete and misleading. MV learning can facilitate a collaborative QoE estimation without direct data sharing, improving prediction accuracy and enabling adaptive optimizations at both the application and network levels to ensure a smooth user experience.

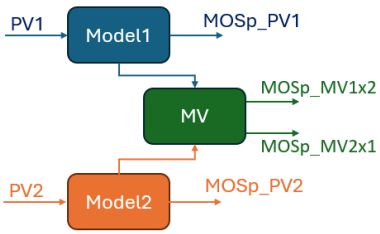


Fig. 1: Multi-view learning for QoE modelling.

Fig. 1 illustrates the application of MV learning for QoE modelling. Let's consider the two partial views (PV1 and PV2) as the data (QoE IFs) collected by different actors, such as ISPs and OTTs, for the same service. Thus, individual models trained on these different data sets would provide different MOS predictions (MOSp_PVx, with $x = 1, 2$) that are both missing some information related to the actual QoE perceived by the end user of the service. Using MV learning, we aim to exploit the data from the two partial views to learn a more comprehensive representation of the collected data and enhance the QoE prediction performance by developing the MOSp_MV model. It is important to highlight that the resulting MOSp_MV model is expected to be more accurate than each PV model, even if only one of the two IFs subsets is used, thanks to the MV training, as clarified in the following.

IV. Model training

This section describes the proposed implementation of MV learning for QoE modeling, which is carried out in three main stages. First, synthetic data sets for training and testing were created (Sect. IV-B) using the TID2008 database as groundwork (Sect. IV-A). Then, the architecture of the NN (Sect. IV-C) and its training strategy (Sect. IV-D) were designed to reflect the principles of the MV approach.

A. TID2008 database

The TID2008 database [10] includes a set of 24 diverse real-world photographs taken from the Kodak database, augmented by one synthetic picture. Each of these 25 reference images is exposed to 17 distinct types of distortion (covering a broad spectrum of common image artifacts, such as additive noise, compression or transmission errors) at four levels of intensity, resulting in a total of 1,700 test images carefully generated to produce various degrees of perceived quality. The database includes an MOS value for each test image, which quantifies the average perceived image quality based on subjective evaluations from over 800 participants. To obtain these judgments, participants performed a series of pairwise comparisons of distorted images relative to a reference picture, selecting the one closer to the reference. The version of the image that looked more like the unperturbed picture would get one additional point. Each distorted picture was inspected nine times, yielding a final vote ranging from 0 to 9. Authors preferred this strategy over the more commonly used technique of assigning absolute ratings because, although requiring more comparisons, it proved less prone to the inaccuracies associated with traditional rating scales, where biases related to scale anchors and fatigue can skew the results. The choice to base our work on this specific database stems from its extensive evaluations detailing the effects of several different image distortion types on the viewer's perceived QoE.

B. Synthetic datasets

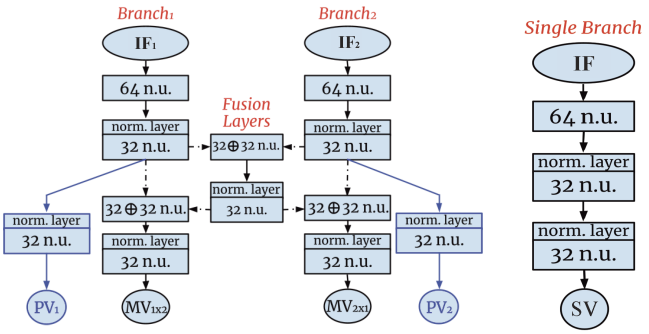
As mentioned in Sect. III, our analysis focuses on the cases where multimedia content and services are affected by more than one distortion at a time, which is not the case for the distorted images in the TID2008 database. This required us to generate a synthetic data set by leveraging the fact that the MOSes in TID2008 are expressed as a function of individual distortion type and intensity level (only one distortion type and level at a time were applied to reference images). This allowed us to assign a weight w_i to each i-th distortion type, based on the MOS variation as a function of the distortion intensity level. These weights were determined using ANOVA (ANalysis Of VAriance), which compares between-groups variance (σ_b) to the within-group variance (σ_w). For each distortion, we calculated the σ_b/σ_w ratio (commonly referred to as *F statistic*) of the MOS to quantify how much the perceived QoE varies with the distortion level[1]. Finally, the 17 values of this metric were normalized ($w_i = F_i / \sum_{i=1}^{17} F_i$) and used as weights to generate the MOS for the synthetic scenarios:

$$MOS_{synt,j} = \sum_{i=1}^{17} w_i \cdot MOS_{i,l}, \tag{1}$$

where $MOS_{synt,j}$ is the j-th synthetic MOS and $MOS_{i,l}$ represents the MOS from TID2008 corresponding to the i-th distortion ($i \in \{1, \ldots, 17\}$), at a randomly drawn distortion level l (with $l \in \{1, 2, 3, 4\}$). To avoid introducing artificial biases into the model, the distortion levels were drawn following a uniform distribution[2].

As evident in Eq. (1), each synthetic MOS (target values) includes the effect of all 17 distortions, which constitute the input fed to the model to be trained. To ensure a large and homogeneous sample and to account for the fact that QoE depends not only on the distortion types and levels but also on the peculiarities of the examined image, we simulated 1,000 scenarios per reference picture, resulting in a total of 25,000 synthetic MOS (making $j \in \{1, \ldots, 25000\}$, in Eq. (1)). This sample size, although it represents a small fraction of the total 25×4^{17} possible scenarios, was chosen as a trade-off to balance computational efficiency with sufficient variability. We acknowledge that our approach assumes independence among impairment types, which is typically not the case in real-life studies. However, in the absence (to our knowledge) of a comprehensive study cross-correlating the effects of the various distortions, this was adopted as the second-best alternative. Moreover, using a data set derived from the already

[1]Python SciPy library provides p-values to assess variability significance.
[2]Distortion levels were drawn using the Python "random.randint" method.



(a) (b)

Fig. 2: Architecture of the proposed FCDNNs. The boxes denote the hidden layers, with the number of n.u. specified inside. a) the MV approach (featuring two independent branches and intermediate fusion layers) and the PV approach (marked in blue). $\mathbf{IF}_1$ and $\mathbf{IF}_2$ represent the sets of n_1 and n_2 IFs, respectively, that serve as inputs to the branches; $PV_{1,2}$ are the outcomes of the partial-views, while $MV_{1\times2}$ and $MV_{2\times1}$ indicate the MV predictions. b) the SV approach, where the entire set **IF** of N IFs is fed to a single branch.

averaged values, rather than from individual participant votes, automatically excluded extreme values from the range of possible prediction outcomes. Nonetheless, since we focused on evaluating the average perceived QoE, extreme values are statistically negligible for our purposes. The generated synthetic MOS values range between 2.86 and 5.69.

C. The neural networks

Fully Connected Deep Neural Networks (FCDNNs) consist of multiple layers, with each neuron in a given layer connected to every neuron in the subsequent layer. This structure enables this kind of network to learn intricate patterns in the data by progressively combining simpler features into high-order representations. FCDNNs are well-suited for QoE prediction tasks due to their ability to model complex, non-linear relationships between input features and output targets [16]. Moreover, FCDNNs excel in processing numerous inputs and learning unified representations that capture the relationships between them. Furthermore, techniques like regularization, dropout, and early stopping help mitigate the risk of overfitting, leading to robust performance even with limited data. The capability of FCDNNs to incorporate intermediate fusion layers also allows them to combine feature representations from different branches, making them highly effective for collaborative QoE modeling, such as the MV approach. In these scenarios, data sets relative to separate views are often fragmented due to concerns about privacy and/or differences in measurement methodologies. By merging only the processed features rather than the raw data, FCDNNs facilitate secure collaborations while maintaining high predictive performance. Finally, adaptability to new scenarios makes FCDNNs particularly well-suited for QoE estimation, as supplementary IFs can be added to previous studies without requiring a complete overhaul of the configuration of the NN.

The architecture of the FCDNN, designed to reflect the MV perspective in this study, is shown in Fig. 2a. It is structured with two independent branches ($Branch_{1,2}$), each assigned a unique and non-overlapping subsample of all the considered IFs ($\mathbf{IF}_{1,2}$). Note that with IFs, we are referring to the 17 distortions discussed in Section IV-A, while the ratio behind the IFs split is detailed in Section IV-D. Both branches process their input through two hidden layers with 64 and 32 neural units (n.u.), respectively. The depth of the NN, as well as the number of neurons that populated each layer, were selected for optimization purposes. On the one hand, each branch passes through an additional layer of 32 n.u. and generates one single QoE prediction. These results correspond to the outcomes of the PV approach, PV_1 and PV_2, and are used as a reference to assess the effectiveness of the MV model. On the other hand, the outputs from the lower hidden layers (32 n.u. each) are concatenated to serve as inputs for a shared fusion branch that starts with a 64 n.u. layer. The fusion branch ends with a second layer composed of 32 n.u. that are integrated back into the last hidden layers of the two independent branches. Each branch then proceeds with two additional hidden layers, with 64 and 32 n.u., before ultimately generating separate QoE predictions, namely $MV_{1\times2}$ and $MV_{2\times1}$, corresponding to the results of the MV approach. Fig. 2b shows the other NN representing the SV approach, which employs a single branch to process the entire IF set. Its design resembles that of an individual branch within the MV approach, having three hidden layers: one with 64 and two with 32 n.u., respectively.

All layers of both NNs are activated by the Rectified Linear Unit ($ReLU(x) = max(0, x)$) activation function. To enhance model stability, intermediate normalization layers ($norm_i$) are incorporated between each i-th layer of the NN:

$$norm_i = \frac{ReLU(\mathbf{x_i}) - \langle ReLU(\mathbf{x_i})\rangle}{\sqrt{var(ReLU(\mathbf{x_i})) + \epsilon}} \cdot \gamma + \beta, \qquad (2)$$

where $ReLU(\mathbf{x_i})$ is the input vector for the i-th layer, $\langle ReLU(\mathbf{x_i})\rangle$ is its mean value, and $var(ReLU(\mathbf{x_i}))$ is its variance; γ and β correspond to the weight and bias, respectively, adjusted during the training process by the Adam optimizer. The artifact $\epsilon = 10^{-5}$ is added to enhance numerical stability. All the parameters of our NNs were determined iteratively as the ones that optimized the performance of our model.

D. Training strategy

The training process for the MV approach involved systematically splitting the $N = 17$ IFs (**IF**) into two mutually exclusive subsets and evaluating all possible configurations of these splits. Specifically, the two branches of the FCDNN were trained with subsets $\mathbf{IF}_1$ and $\mathbf{IF}_2$ of n_1 and n_2 IFs as input, where $n_1 + n_2 = N$. The aim was to train the model for all distinct ways in which the IFs could be partitioned into two non-overlapping groups, covering every possible pairing of n_1 and n_2. For each configuration, the two subsets were generated by selecting all possible combinations of n_1 elements from

TABLE I: Performance of the MV and PV approaches, averaged across each IF split, compared to that of the SV approach (bottom row). Metrics: the mean (Avg), median and third quartile (Q3) of the absolute difference between the predictions and the validation values (|diff|); the proportion of predictions deviating by at least 0.3 (%|diff|≥0.3); RMSE of the diff distribution.

Split N.	IFs Split	Approach	Avg |diff|	Median |diff|	Q3 |diff|	%|diff|≥0.3	RMSE (diff)
1	$n_1 = 16$	MV_{1x2} (PV_1)	0.117 (0.145)	0.096 (0.113)	0.169 (0.202)	4.64 (10.54)	0.148 (0.194)
	$n_2 = 1$	MV_{2x1} (PV_2)	0.118 (0.411)	0.096 (0.377)	0.170 (0.594)	4.80 (60.26)	0.149 (0.492)
2	$n_1 = 15$	MV_{1x2} (PV_1)	0.116 (0.167)	0.095 (0.127)	0.167 (0.228)	4.45 (14.92)	0.147 (0.225)
	$n_2 = 2$	MV_{2x1} (PV_2)	0.116 (0.396)	0.096 (0.361)	0.168 (0.574)	4.52 (58.14)	0.148 (0.477)
3	$n_1 = 14$	MV_{1x2} (PV_1)	0.116 (0.189)	0.095 (0.142)	0.167 (0.259)	4.74 (19.45)	0.148 (0.254)
	$n_2 = 3$	MV_{2x1} (PV_2)	0.117 (0.382)	0.096 (0.344)	0.168 (0.555)	4.87 (56.11)	0.149 (0.464)
4	$n_1 = 13$	MV_{1x2} (PV_1)	0.117 (0.211)	0.096 (0.159)	0.169 (0.292)	4.70 (23.94)	0.148 (0.281)
	$n_2 = 4$	MV_{2x1} (PV_2)	0.117 (0.370)	0.096 (0.329)	0.170 (0.539)	4.82 (53.99)	0.149 (0.452)
5	$n_1 = 12$	MV_{1x2} (PV_1)	0.116 (0.230)	0.096 (0.177)	0.169 (0.323)	4.43 (27.97)	0.148 (0.304)
	$n_2 = 5$	MV_{2x1} (PV_2)	0.117 (0.353)	0.097 (0.308)	0.169 (0.514)	4.54 (51.08)	0.148 (0.435)
6	$n_1 = 11$	MV_{1x2} (PV_1)	0.116 (0.250)	0.096 (0.194)	0.168 (0.356)	4.43 (31.88)	0.147 (0.326)
	$n_2 = 6$	MV_{2x1} (PV_2)	0.117 (0.337)	0.096 (0.290)	0.168 (0.493)	4.51 (48.47)	0.148 (0.418)
7	$n_1 = 10$	MV_{1x2} (PV_1)	0.116 (0.268)	0.096 (0.212)	0.168 (0.385)	4.46 (35.45)	0.147 (0.346)
	$n_2 = 7$	MV_{2x1} (PV_2)	0.117 (0.320)	0.096 (0.268)	0.168 (0.468)	4.51 (45.27)	0.148 (0.401)
8	$n_1 = 9$	MV_{1x2} (PV_1)	0.115 (0.285)	0.095 (0.230)	0.167 (0.415)	4.26 (38.88)	0.146 (0.364)
	$n_2 = 8$	MV_{2x1} (PV_2)	0.116 (0.302)	0.096 (0.249)	0.167 (0.442)	4.27 (42.15)	0.146 (0.382)
		SV	0.117	0.096	0.168	4.81	0.148

the total N, with the remaining n_2 elements forming the complementary subset. The number of unique combinations for a given split is determined by the binomial coefficient $C_{n_1,n_2} = N!/(n_1! \cdot n_2!)$, where n_1 was varied from 16 to 9, and consequently n_2 varied from 1 to 8. The number of configurations explored for each value of n_1 is: $C_{(16,1)} = 17$; $C_{(15,2)} = 136$; $C_{(14,3)} = 680$; $C_{(13,4)} = 2,380$; $C_{(12,5)} = 6,188$; $C_{(11,6)} = 12,376$; $C_{(10,7)} = 19,448$; $C_{(9,8)} = 24,310$. Therefore, the total number of models trained across all configurations is given by $C_{tot} = \sum_{n_1=9}^{16} C_{n_1,n_2} = 65,535$. This exhaustive combinatorial approach ensured that the entire space of feature partitions was explored, accounting for every possible way the IFs can be split into two groups and allowing a detailed analysis of how the composition of subsets influences the predictive performance of the FCDNN. The NN corresponding to the SV approach was trained only once. To ensure a reliable performance evaluation, the data set was partitioned into 70% for training and 30% for validation. This split balanced adequate training data and a large validation set for accurate model evaluation. The Mean Squared Error (MSE) loss function was selected during this training stage.

The hyperparameters used during the model training process were carefully selected to promote stable and efficient training while preventing overfitting and balancing computational efficiency with model generalization capabilities. A batch size of 2048 was chosen to leverage computational advantages, enabling faster training convergence and more stable gradient estimates. A learning rate of $6.4 \cdot 10^{-4}$ was selected to ensure a stable and smooth convergence of the training process, avoiding oscillations or divergence and improving stability despite potentially increasing training time. A maximum of 1,000 epochs was allocated to allow ample time for the network to learn patterns in the data. However, the actual training was stopped earlier if validation performance did not improve for 100 consecutive epochs to prevent overfitting.

V. Model assessment

The predictive performance of our model was evaluated by measuring the difference between its predictions and the target values at two different levels of granularity. First, the MV approach was analyzed on a split-by-split basis to determine which configuration yielded the best performance, on average, compared to the SV approach and exhibited the largest improvement over the corresponding PV approach. Subsequently, we identified the combinations of IFs that delivered the highest performance within the best-performing split.

Table I shows that the MV approach achieved results comparable to the SV approach (if not better than that, in some cases) across all IFs splits and evaluation metrics, including: the average (Avg), median, and third quartile (Q3) values of the absolute difference (|diff|) between the predictions and the validation targets; the percentage of absolute differences exceeding an arbitrarily chosen threshold (%|diff| ≥ threshold)[3]; and the Root Mean Squared Error (RMSE) of the distribution[4]. Each row represents the performance of a single split, with the SV approach results reported in the bottom row. The nomenclatures "1x2" and "2x1" refer to the MV branches that, before the fusion layer, received n_1 and n_2 inputs, respectively.

The most favorable results, compared to the SV approach, were obtained with "split 8" ($n_1 = 9$, $n_2 = 8$), which suggests that the model achieves its highest predictive power when the input data set is approximately equally divided between the two separate views. This might be connected to the fact that the two branches, when processing a similar number of IFs, produce feature representations that are more balanced and coherent before being merged in the fusion layer. As a result, the integrated information may be more consistent, reducing

[3]The threshold was set to 0.3 since such a deviation defines the range within which a MOS can be rounded to the nearest next/previous half-integer.

[4]To be noticed that the RMSE is the only metric to be calculated over the distribution of *diff*, rather than on its absolute values.

TABLE II: Combinations of input IFs that resulted in the highest performance gain for the MV approach, compared to SV, evaluated across split 8. The "Standalone" results concern the combinations yielding the highest improvement for each metric, whereas the "Overall" results regard the combination leading to the greatest overall enhancement across all metrics.

	Metric	SV	MV_{1x2}			MV_{2x1}			IF Combination
			MV_{1x2}	PV_1	Perf. gain	MV_{2x1}	PV_2	Perf. gain	
Standalone	Avg |diff|	0.117	0.109	0.210	6.64%	0.109	0.402	6.52%	[0, 5, 6, 7, 9, 11, 12, 13, 14], [1, 2, 3, 4, 8, 10, 15, 16]
	Median |diff|	0.096	0.089	0.387	6.73%	0.090	0.181	6.15%	[0, 2, 4, 5, 6, 8, 13, 14, 16], [1, 3, 7, 9, 10, 11, 12, 15]
	Q3 |diff|	0.168	0.155	0.307	7.75%	0.157	0.549	6.71%	[0, 5, 6, 7, 9, 11, 12, 13, 14], [1, 2, 3, 4, 8, 10, 15, 16]
	%|diff|≥0.3	4.810	3.010	63.030	159.73%	3.170	16.040	151.68%	[2, 4, 5, 6, 7, 10, 11, 13, 15], [0, 1, 3, 8, 9, 12, 14, 16]
	RMSE (diff)	0.148	0.137	0.262	107.95%	0.138	0.456	107.43%	[0, 5, 6, 7, 9, 11, 12, 13, 14], [1, 2, 3, 4, 8, 10, 15, 16]
Overall	Avg |diff|	0.117	0.109	0.400	6.11%	0.109	0.211	6.54%	[0, 2, 4, 5, 6, 8, 13, 14, 16], [1, 3, 7, 9, 10, 11, 12, 15]
	Median |diff|	0.096	0.089	0.387	6.73%	0.090	0.181	6.15%	
	Q3 |diff|	0.168	0.156	0.546	7.03%	0.156	0.306	7.01%	
	%|diff|≥0.3	4.810	3.200	64.670	150.42%	3.370	26.040	142.69%	
	RMSE (diff)	0.148	0.139	0.454	106.50%	0.138	0.262	106.74%	

discrepancies between the two branches and enhancing the generalizability of the model. All PV configurations, whose performance values are shown within brackets, significantly benefited from the additional information provided by the fusion layer. The branch that experienced the greatest enhancement due to the inclusion of the merging layer was the one initially receiving only a single IF as input ("split 1"), whose performance became comparable to that of SV after the implementation of the fusion layer. However, this outcome is primarily attributed to the weak performance of the PV approach in that case, rather than to the merits of the MV approach. Nevertheless, although the advantage of the MV approach diminishes as the number of IFs in the input set approaches the total data set size, a consistent improvement is observed across all evaluation metrics.

Table II shows the combinations of IFs that yielded the largest improvements compared to the SV approach for MV_{1x2} and MV_{2x1} within "split 8". To quantify these performance improvements, different comparison strategies were employed depending on the nature of each metric. Specifically, for Avg |diff|, Median |diff|, and Q3 |diff|, the gain was expressed as a relative variation using the ratio $100 \cdot (X_{SV} - X_{MV})/X_{SV}$ (where X represents the considered metric), since these metrics are point estimates. Conversely, for %|diff| $\geq$ 0.3 and RMSE, a direct ratio $100 \cdot X_{MV}/X_{SV}$ was used, as these metrics describe the scale of variation rather than a central reference point. Moreover, the 'Standalone' results refer to the IF combinations maximizing the improvements for each single metric, whereas the 'Overall' results concern the IF combination that maximized the product of the improvements across all metrics. Among the 24,310 IF combinations tested for "split 8", the one achieving the largest overall performance gains is $[0, 2, 4, 5, 6, 8, 13, 14, 16], [1, 3, 7, 9, 10, 11, 12, 15]$.

These results confirm that the fusion of PV complementary representations from distinct subsets of IFs not only compensates for individual branch weaknesses but also produces more generalizable predictions than both the PV and SV approaches. Thus, leveraging diverse perspectives through an MV strategy is a promising direction for QoE modeling. However, it is important to highlight that these results are achieved from training on a synthetic dataset in a scenario specifically designed to isolate and evaluate the contribution of the MV approach. Moreover, this dataset relies on a simplified, linear additive model of MOS computation subjected to multiple distortions, which may not directly reflect ground-truth MOS subjected to interacting distortions of real-world content. Nonetheless, the results offer valuable insight into the potential applicability of MV learning in such contexts.

VI. Conclusions

In this paper, we investigated the potential of MV learning for QoE modeling and prediction within a representative yet simplified scenario. Our results demonstrate that the proposed MV framework achieves comparable or superior predictive performance compared to SV, regardless of the chosen partitioning scheme. Specifically, the best-performing scenario is achieved when the input data set is approximately equally divided between the two separate views in terms of the number of features, which led to an improvement of 6.15% – 142.69% across most evaluation metrics, compared to the SV approach.

Overall, our findings highlight the potential of the MV approach as a powerful tool for QoE estimation compared to PV and SV, especially in privacy-sensitive scenarios. Nonetheless, it is worth noting that these outcomes are derived from training on a synthetic dataset based on a simplified, linear additive model for computing synthetic MOS values, weighting the influence of multiple co-occurring distortions. This setup limits the direct generalizability of the quantitative findings to real-world QoE scenarios.

Future work will focus on validating the MV approach using datasets that feature real-world content affected by multiple, naturally co-occurring and interacting distortions, with ground truth MOS collected with subjective quality assessments.

References

[1] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A survey on parametric QoE estimation for popular services," *Journal of Network and Computer Applications*, vol. 77, pp. 1–17, 2017.

[2] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 2s, May 2018.

[3] Y. Li, M. Yang, and Z. Zhang, "A Survey of Multi-View Representation Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2019.

[4] Z. Yu, Z. Dong, C. Yu, K. Yang, Z. Fan, and C. L. P. Chen, "A review on multi-view learning," *Frontiers of Computer Science*, vol. 19, no. 7, p. 197334, 2024.

[5] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[6] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3539–3553, 2023.

[7] Y. Wang, Y. Xiao, J. Lu, B. Tan, Z. Cao, Z. Zhang, and J. T. Zhou, "Discriminative Multi-View Dynamic Image Fusion for Cross-View 3-D Action Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5332–5345, 2022.

[8] J. Chen, Y. Wang, and Y. Y. Tang, "Person Re-identification by Exploiting Spatio-Temporal Cues and Multi-view Metric Learning," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 998–1002, 2016.

[9] M. Hamidi, S. Porcu, A. Floris, and L. Atzori, "Towards the Application of Multi-view Learning in Quality of Experience Collaborative Modelling," in *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, 2024, pp. 286–292.

[10] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "TID2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[11] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, ser. COLT' 98. New York, NY, USA: Association for Computing Machinery, 1998, p. 92–100.

[12] T. Scheffer, "Email answering assistance by semi-supervised text classification," *Intell. Data Anal.*, vol. 8, no. 5, p. 481–493, Oct. 2004.

[13] J. Chen, L. Yang, L. Tan, and R. Xu, "Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition," *Pattern Recognition*, vol. 129, p. 108753, 2022.

[14] S. Ickin, M. Fiedler, and K. Vandikas, "QoE Modeling on Split Features with Distributed Deep Learning," *Network*, vol. 1, no. 2, pp. 165–190, 2021.

[15] S. Ickin, D. Roeland, and G. Hall, "Independent Split Model Inference at Operator Network for Network Performance Estimation," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, 2023, pp. 1–9.

[16] X. Liu, G. Chuai, X. Wang, Z. Xu, and W. Gao, "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.