CB-FL: Cluster-Based Federated Learning applied to Quality of Experience modelling

Simone Porcu[1,2], 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
{simone.porcu, alessandro.floris84, l.atzori}@unica.it

***Abstract*—The Federated Learning (FL) approach can be exploited to build a solution to data sparsity and privacy protection issues (e.g., utilization of user-sensitive data) in Quality of Experience (QoE) modelling. In this paper, we investigate whether it is possible to obtain improvements in FL-based inference by grouping data sources to build separate inference systems. To this, we adopted an experimental based approach: firstly, we identified different clusters of users, from a public QoE dataset, based on user-related QoE influence factors and the distributions of the quality rating scores provided by the users; secondly, we developed a Cluster-Based FL QoE predictor and conducted experimental tests to compare the QoE prediction performance with that obtained by a centralised learning approach and a standard FL approach. The obtained results show that the proposed approach achieved the best QoE prediction performance (in terms of accuracy, precision, recall, and F1-Score), followed respectively by the standard FL and the centralised approach.**

***Index Terms*—Quality of Experience, Federated Learning, QoE estimation, Neural network, Collaborative Learning, Clustering.**

I. Introduction

Accurate modelling of the user Quality of Experience (QoE) is becoming more and more important to deliver successful multimedia services over the network. The resulting models are used in resource management operations to reduce customers churn and increase their engagement, which is nowadays extremely important as more and more users are accessing to bandwidth-hungry multimedia applications over limited network resources [1]. Accuracy in the modelling is obtained by creating *personalized* QoE models [2], which means identifying as many influencing factors as possible, especially in the human domain, that highlight very different reactions in the perceived quality when the same network and application settings are applied.

However, realizing personal QoE models from a multi-dimensional perspective requires adequate observable user's data, which is difficult to obtain from the individual user. Therefore, Collaborative Learning (CL) approaches are commonly used to avoid the data sparsity issues in Artificial Intelligence (AI) based modelling. Typically, with CL, data from different users is observed and shared to collect datasets of significant size and with adequate level of information. However, such an approach introduces privacy-protection issues, as the exchanged data may contain user-sensitive data. Thus, the Federated Learning (FL) approach can be used in CL for QoE modeling in a decentralized setting. Indeed, with FL each user's raw data is stored locally in the user's device and not exchanged or transferred; a partial training is performed in the device and the obtained weights from the trained network are shared to achieve the aggregated learning objective [3].

Limited efforts have been devoted to study the application of FL to QoE modelling till now. In [4], two CL approaches (i.e., Round-Robin Learning and FL) are proposed to share trained model weights of isolated datasets to overcome the small QoE data lakes issue while preserving privacy at the same time. The accuracy of the resulting CL QoE models are compared with that obtained by Isolated Learning (IL) QoE models, which do not share any data. Evident improvements have been obtained; however, only binary prediction models were implemented. An FL architecture is also proposed in [2] with the aim to enhance the performance of a personalized QoE model, i.e., a model that considers personalized characteristics (e.g., user, device, and context factors) to predict the user's QoE. However, this work does not provide a clear comparison with a centralized CL approach. The FL was the key for sharing these user-related data to resolve the data sparsity issue while protecting the user privacy.

With the objective to achieve personalized QoE models, more and more features are considered in FL with reference to the users context and profile as well as network and system settings. This is done in response to an ever increasing heterogeneity of service utilization setup and quality demands from the users. However, this increases the sparsity-issues of the collected datasets which results into QoE models that are not able to capture the attitude of users that have dissimilar behaviors with respect to average profiles, represented by the Mean Opinion Score (MOS); in other words, it results that merging datasets that include heterogeneous clusters of users reduces the representativeness of unique users which cannot

This work has been partially funded by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) with the PON "Ricerca e Innovazione" 2014-2020 (PON R&I) Action IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione", assigned with D.M. 1062 on 10.08.2021.

DOI: 10.1109/SITIS57111.2022.00093

be embodied in the inference model. For this reason, it may be beneficial to create separate models for distinctive clusters of users (and relevant data) so as to maintain their peculiarities.

Therefore, the research question we tackle in this paper is: *Is it possible to obtain improvements in FL-based inference by grouping data sources (users) to build separate inference systems?* In this paper we adopt a purely experimental-based approach. Specifically, we performed preliminary data analyses on a public QoE dataset [5] to identify different clusters of users based on user-related QoE influence factors and the distribution of the single provided quality rating scores. Then, we developed a Cluster-Based FL (CB-FL) QoE predictor to estimate the user's perceived QoE. We compared the proposed method with the centralised learning approach and the standard FL approach to answer to the above formulated question. The performance of the implemented methods was compared in terms of mean accuracy, precision, recall, and F1-Score metrics concerning the prediction of the QoE scores.

The paper is structured as follows. Section II discusses the Related work. In Section III, we present the considered scenario and the proposed Cluster-Based FL approach as well as the conducted data analysis and preprocessing operations. The experimental results are described in Section IV, and finally Section V concludes the paper.

II. Related Work

The most common approach used to create AI-based QoE prediction models is the centralised learning approach, which consists in transferring the training data to a central node. A centralised AI QoE predictor for video streaming services is proposed in [6], where the authors focused on the prediction of the MOS based on the network-related influence factors. Their implemented multilayer feedforward deep learning neural network (DNN) reached a mean prediction error lower than 0.25 and a mean square error of 0.15. The authors in [7], implemented a centralised Convolutional Neural Network (CNN) for real-time QoE prediction in video streaming service, providing also the optimal hyperparameters for the model. Following the centralised approach, the authors in [8] trained a DNN to provide a general model to predict the QoE. In [9], the authors propose a two-step QoE modelling approach to understand the correlation between subjective preferences and services. Therefore, subjective preferences are used to create a model to understand the best service conditions for the single user. In [10], the authors propose a continuous-time video QoE predictor that effectively captures the effects of a variety of QoE-influencing factors, and that is able to accurately predict viewers' instantaneous QoE. A continuous time prediction QoE prediction model is also proposed in [11], which relies on simple, but highly descriptive "QoE-aware" inputs: objective video quality assessment (VQA), playback status information, and QoE memory descriptors. These inputs are continuously measured on videos and continuously fed into a nonlinear prediction engine expressed as a single hidden layer neural network (NN). In order to predict the impact of video impairment events such as bitrate drop and rebuffering on QoE, the authors in [12] have proposed a continuous QoE prediction model. The inputs of the prediction model consist of frame quality, rebuffering event state, and the vector characterizing memory effect. The proposed model uses a block-structured nonlinear Hammerstein-Wiener model. The drawbacks of the aforementioned centralised-based approaches are that they do not consider: i) the privacy concerns of the sensitive user data that is required to be shared; ii) the bandwidth that a continuous data transmission consumes; iii) and the computational complexity needed at the servers side to manage all the data transfers from the users.

To overcome the data sharing privacy problem, the research introduced the FL architecture. FL is defined as "a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider: Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective" [3]. The FL is used in many application scenarios, such as proactive caching of multimedia contents [13], protecting the confidentiality of local updates in 5G networks [14], and optimizing mobile edge computing, caching and communication by intelligently utilizing the collaboration among devices [15]. However, the application of the FL approaches to the QoE modelling is quite limited in the literature. In [4], two CL approaches (i.e., Round-Robin Learning (RRL) and FL) are used to share trained model weights of isolated datasets to overcome the small QoE data lakes issue by preserving privacy. The accuracy of the CL QoE models was compared with that obtained by Isolated Learning (IL) QoE models, which do not share any data. Experimental results highlight that CL approaches have the potential to outperform IL approaches, while also protecting privacy issues. The users in the dataset were randomly divided in three groups before performance evaluation. However, since the used dataset was small, the MOS scores were divided in 2 classes and binary prediction models were implemented. An FL architecture is also used in [2] with the aim to enhance the performance of a personalized QoE model, i.e., a model that considers personalized characteristics (e.g., user, device, and context factors) to predict the user's QoE. The FL was the key for sharing the user-related data to resolve the data sparsity issue while protecting the user privacy. However, it is not clear how the clustering analysis was applied to identify different clusters of users based on the data distribution of video watching hours, which is a parameter difficult to be measured in a real application scenario.

In this work, we consider both the centralized approach and the FL approach. In particular, we aim to investigate whether is it possible to obtain improvements in FL-based inference by grouping data sources (users) to build separate inference systems. We estimate the QoE on the 5-level ACR scale and not on the binary MOS scale as in [4].

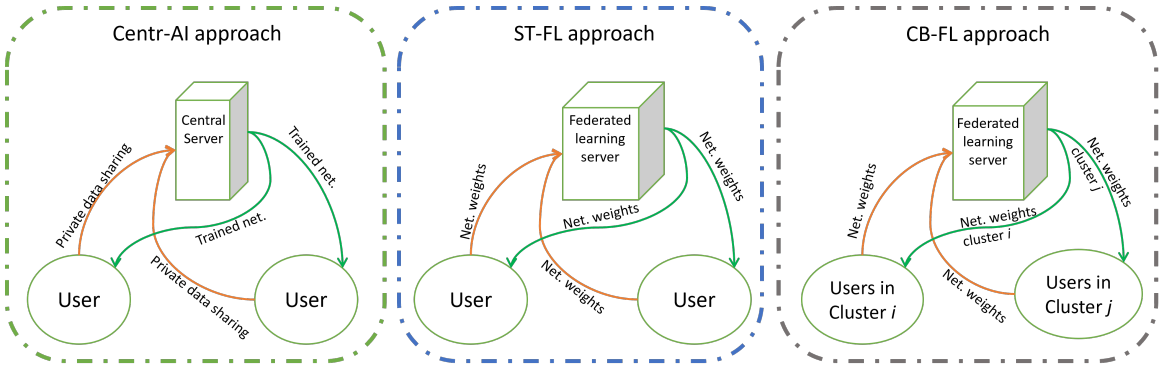

Fig. 1. The three considered Collaborative Learning approaches.

III. Cluster-Based FL

In Section III-A, we present the considered reference scenario and the proposed Cluster-based FL approach. Section III-B describes the selected dataset, whereas Section III-C and Section III-D illustrate respectively the performed data analysis and data preprocessing operations.

A. Reference scenario

In the considered reference scenario n users utilize a multimedia service over the network under varying network and application parameters, and they provide a feedback about the perceived QoE every time the system asks them to do so. The collected feedback is used to build a knowledge on the user quality perception to optimize the service deployment and management in the future. In particular, we focused on the creation of QoE models based on the users' feedback and observed parameters using the three different approaches that are shown in Fig. 1, i.e., the centralised AI (Centr-AI), the standard FL (ST-FL), and the proposed CB-FL approaches.

In the Centr-AI approach, the perceived quality and the objective impact factors of each user are shared with a central server, which is in charge of training the model and to send it back to each user. The model is generated with data collected from different users, which may belong to different clusters on the basis of their reaction to the service configuration and performance parameters. Using data from quite different clusters may negatively impact on the prediction capabilities of the inference module. Another major drawback of the Centr-AI approach is that the users need to transfer all the data to the central server, which may lead to privacy issues in case user-sensitive data is involved.

To overcome the privacy issue of the Centr-AI approach, due to the need to share all the user's data, we consider the ST-FL approach. With this approach, the user's data is stored locally and not exchanged or transferred; instead, the model is trained in the user's device and only the weights of the trained NN are shared with the FL server. The FL server, in turn, creates a combined NN that will work for all users, and shares the weights of this network with each user. However, this approach can work efficiently only if there are enough users participating in the training stage. Also, similarly to the Centr-AI approach, the prediction model implemented by the ST-FL approach may suffer from data belonging to quite different clusters, which may prevent the prediction model to focus on possible differences in quality perception of clusters of users.

Finally, with the proposed CB-FL approach, we aim to preserve the user's privacy (thanks to the FL architecture) as well as to provide enhanced QoE prediction models based on the grouping of user data sources to build separate inference systems. Indeed, as stated before, the QoE models provided by the Centr-AI and ST-FL approaches reflect the general behavior of dissimilar clusters of users participating in the model training. However, since the QoE is subjective, it may happen that different users perceive different QoE, even when same network and application parameters are measured, because of, for example, different background, different quality expectations, different context of use or different used device. For these reasons, with the proposed CB-FL approach we aim to perform preliminary data analysis to identify user clusters (in terms of user- and/or context-related data) with similar characteristic in terms of quality perceptions. Then, with the FL architecture, specific QoE prediction models can be implemented for different clusters so as to provide enhanced QoE prediction to users with different expectations. As shown in Fig. 1, with the CB-FL approach the FL server receives the weights from the users belonging to a specific cluster and creates a combined NN that will work for all users of that cluster, by sharing the weights of this network with them. Our research study aims at evaluating whether the QoE models created for specific clusters of users would estimate the QoE with major prediction accuracy than that achieved by the generic models obtained with the Central-AI or the ST-FL approaches.

B. Dataset

To perform experimental evaluation and compare the three considered approaches, we based on the Poqemon-QoE-Dataset [5]. The dataset includes 1560 samples of subjective ACR (Absolute Category Rating) scores provided by 181 users concerning the QoE of multimedia contents watched on the VLC media player under several test conditions.

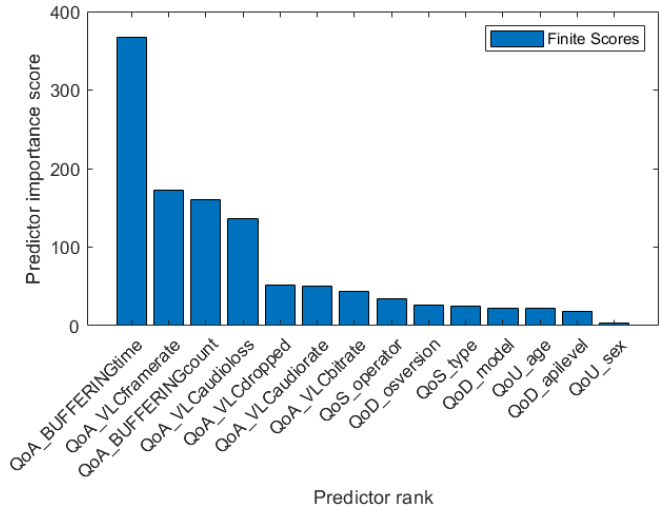

Fig. 2. Univariate feature ranking results computed using the chi-square test on the complete dataset.

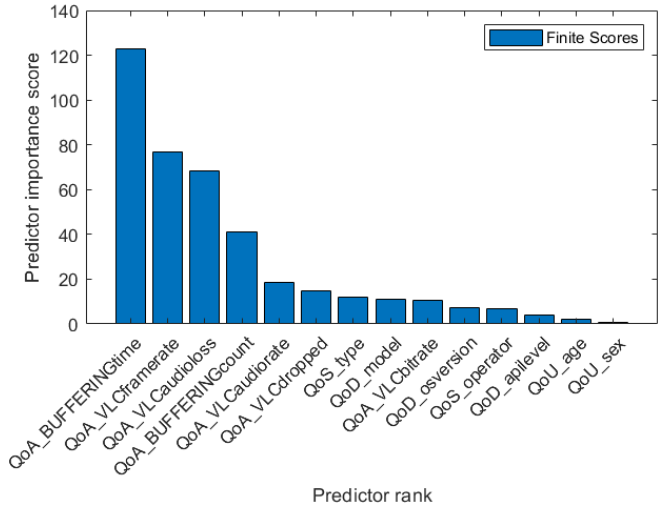

Fig. 3. Univariate feature ranking results computed using the chi-square test on the cluster 14-25.

The observed QoE influence factors (IFs) were 16, classified in 4 categories:

- Quality of Application (QoA): VLC resolution, VLC bitrate, VLC Framerate, VLC dropped frames, VLC audio rate, VLC audio loss, number of buffering times, and the sum of buffering seconds.
- Quality of Service (QoS): network type (i.e., Edge, UMTS, HSPA, HSPAP, LTE), and network operator (i.e. SFR, Bouyegues, Orange, free).
- Quality of Device (QoD): device model, operating system and used API level.
- Quality of User (QoU): age, sex, and study level of the participants.

The participants provided a quality feedback, using the 5-points ACR scale (1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) to evaluate the overall QoE.

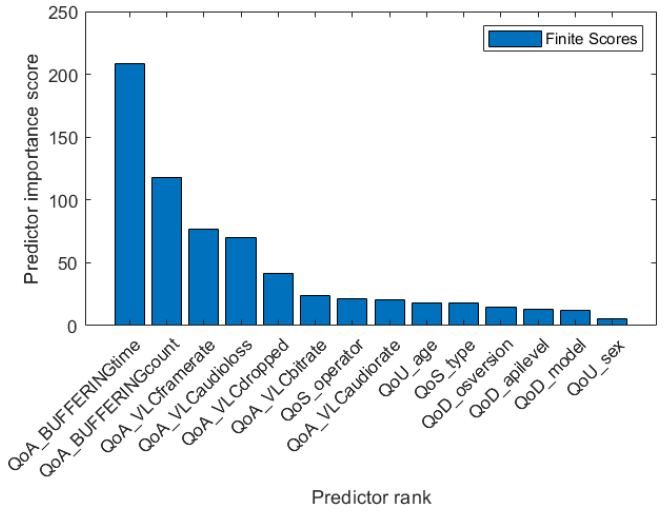

Fig. 4. Univariate feature ranking results computed using the chi-square test on the cluster 26-32.

C. Preliminary data analysis

Since the data of the Poqemon-QoE-Dataset was collected in a real usage scenario and not in a laboratory environment, not every participant rated the QoE for each combination of IFs. For this reason, unbalanced distributions of ACR scores were obtained. Therefore, we have first computed the K-means clustering algorithm on the ACR data to identify clusters including balanced and complete distributions of ACR scores, i.e., with a comparable number of samples for each quality rating scale. We set as the constraint that these ACR distributions should belong to clusters of users categorized on the basis of a user-related attribute, independent from network and application parameters, which would be potentially useful to clusterize the users in different groups. We then computed the clustering by using respectively the age, the sex, and the study level of the participants as the constraint. The age attribute provided the best result in terms of cluster separation, and we eventually found that the optimal number of clusters was 2, the first cluster containing the data of users from 14 to 25 years old, and the second cluster containing the data of users from 26 to 32 years old.

Then, to verify that the 2 clusters had the potential to effectively reflect different groups of users with different quality perceptions, we computed the chi-square test on the data. Precisely, we computed the chi-square test on:

- The complete dataset (Fig. 2).
- The 14-25 cluster (Fig. 3).
- The 26-32 cluster (Fig. 4).

The aim of this test was to investigate whether the different IFs considered in the subjective test achieved different significance scores with respect to the perceived quality (ACR scores). The predictor importance score is computed as $-log(p)$, where p is the p-value obtained from the test statistics. The lower the p-value the stronger is the correlation between the IF and the ACR score and therefore the greater is the impact of that IF on

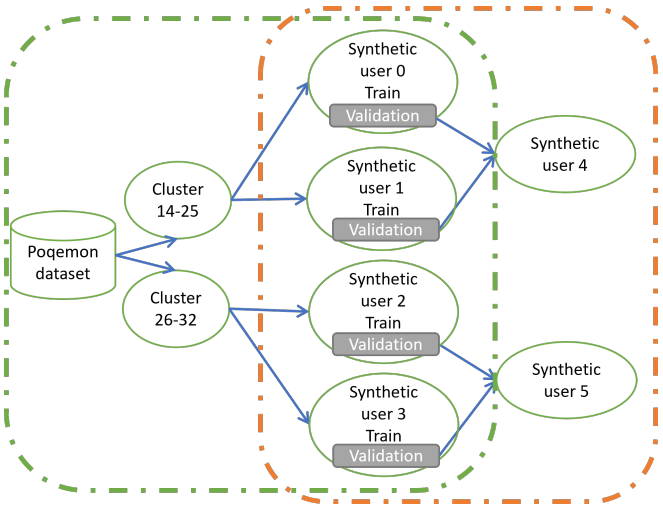

Fig. 5. Data preprocessing.

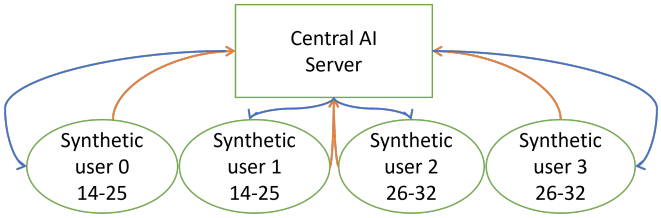

Fig. 6. Centralised AI server approach.

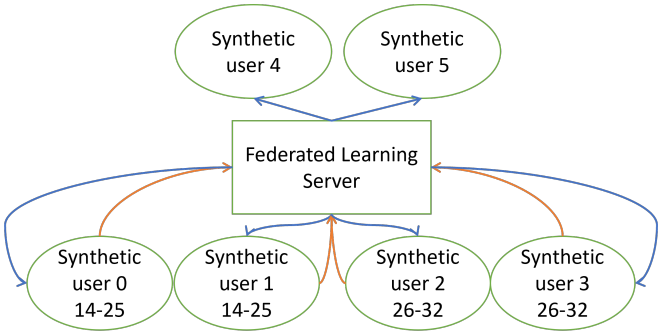

Fig. 7. Standard Federated Learning approach.

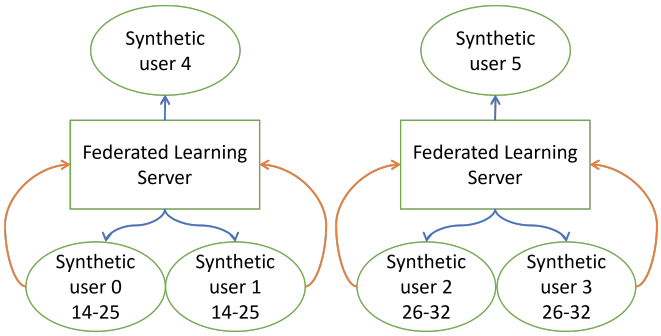

Fig. 8. Cluster-Based Federated Learning approach.

the perceived QoE. Note that the QoU study level and the QoA video resolution IFs are not shown because the obtained p-value was not significant (greater than the threshold of 0.001). From Figs. 2-4, it can be seen that while the buffering time is the most significant IF for all cases, the significance order of the other IFs is different for the 2 clusters. This means that the obtained clustering result would potentially reflect 2 groups of users that perceive the QoE differently. On the basis of this result, we then assumed that training a NN for each different user cluster would lead to better QoE prediction than training a single NN on the complete dataset. In Section IV, we perform experimental tests to verify this assumption.

D. Data preprocessing

Before conducting the experimental tests, data preprocessing operations were needed to prepare the training and validation datasets that will be the input of the NNs of the considered modelling approaches. This step was needed because the FL requires an important amount of data to work efficiently and provide accurate prediction results. Thus, we decided to create two synthetic users from each of the two clusters' data described in Section III-C. The aim of the synthetic user is to *simulate* a single user with particular quality expectations that provided a relevant number of QoE evaluations, using the full ACR scale, with regard to different test conditions. Since such a single user with these characteristics does not exist in the used dataset, we created it artificially by grouping the QoE evaluations of the users included in the considered clusters. Fig. 5 illustrates the data preprocessing operations. Thus, each cluster's data was divided into two synthetic users (Synthetic user 0 and Synthetic user 1 for Cluster 14-25 and Synthetic user 2 and Synthetic user 3 for Cluster 26-32) containing the full scale of ACR rates. This operation was performed selecting randomly the scores in a way to have balanced distributions in the full range for the resulting synthetic users. Each synthetic user includes enough data to participate at the FL training stage, so it has a training set and a validation set, as highlighted in the green box of Fig. 5. For instance, the Cluster 14-25 includes 74 users with a total of 753 quality evaluations, which were divided to obtain two complete (in terms of the utilization of the full ACR rating scale) subsets for creating respectively the "Synthetic user 0" (376 quality evaluations) and the "Synthetic user 1" (377 quality evaluations). A synthetic user cannot contain the data of an user already contained by another synthetic user.

Furthermore, to further evaluate the performance of the NNs trained with the data of these synthetic users, we created an additional synthetic user (for each of the clusters) composed of the validation subsets of the respective synthetic users. These additional synthetic users are called "Synthetic user 4" and "Synthetic user 5" and are highlighted in Fig. 5 (orange box).

IV. Experimental Results

In Section IV-A, we describe the conducted experiment whereas Section IV-B discusses the achieved experiment results.

A. Experiment description

The objective of the experiment was to compare the QoE prediction performance of the three approaches described in Section III, i.e., the Centr-AI approach, the ST-FL approach, and the proposed CB-FL approach.

As shown in Fig. 6, the Centr-AI approach requires that all the synthetic users (0, 1, 2, and 3) share their complete dataset with the central AI server. The server merges the four datasets (respectively the training set and the validation set) and trains the NN doing a 5-fold cross validation with a $70\%/30\%$ training/validation rate to compute the final results in terms of mean accuracy, precision, recall and F1-score.

Instead, both the ST-FL and CB-FL approaches follow a different configuration based on the data preprocessing described in Section III-D and illustrated in Fig. 5. Indeed, for these two approaches, the "Synthetic user 4" and "Synthetic user 5" are used as the validation sets to train the NN. Therefore during the FL training stage, the single synthetic users (0, 1, 2, and 3) trains the NN doing internally (based on their own dataset splitting) a 5-fold cross validation with a $70\%/30\%$ training/validation rate to compute the final results in terms of mean accuracy, precision, recall and F1-score. Then, the obtained NNs are validated using the "Synthetic user 4" and "Synthetic user 5". The difference between the ST-FL and the CB-FL regards the training and validation of the NNs. The ST-FL uses the union of "Synthetic user 4" dataset and "Synthetic user 5" dataset as the validation set to train a single NN (see Fig. 7) for all Synthetic users (0, 1, 2, and 3). The CB-FL approach trains two different NNs based on the age-based clustering results (see Fig. 8), i.e., one NN for the synthetic users concerning the cluster 14-25 (validated on "Synthetic user 4") and another NN for the synthetic users concerning the cluster 26-32 (validated on "Synthetic user 5").

The NN considered for the three approaches was a DNN composed of 4 dense layers, each activated by the Rectified Linear Unit (ReLu) activation function and a Softmax Layer for the ACR score prediction. To implement the FL infrastructure, we used Pytorch and Flower frameworks. These two frameworks are compatible and allow the development of NNs and FL infrastructures thanks to the well-described libraries.

The implemented NN has a "lightweight" configuration since this experiment wants to provide a usable QoE predictor for real environments that could involve smartphones or tablet devices powered by a single battery pack. Thus, a deeper NN could accelerate the device battery drain if it not connected to the electric network. To obtain comparable results we trained the DNN with 200 round of training that required 3 minutes to complete the training session.

B. Results

Tables I, II and III show respectively the QoE prediction performance of the Centr-AI approach, the ST-FL approach and the CB-FL approach, in terms of mean accuracy, and precision, recall, and F1-score of the single ACR scores. The Centr-AI approach achieved a mean accuracy of 0.75. However, from the precision, recall and F1-score results it can

TABLE I
QOE PREDICTION PERFORMANCE OF THE CENTR-AI APPROACH.

Synthetic User	Metric	ACR				
		1	2	3	4	5
All	Mean Accuracy	0.75				
	Precision	0.67	0.50	0.57	0.82	0.85
	Recall	0.69	0.49	0.52	0.86	0.82
	F1-Score	0.68	0.50	0.54	0.84	0.75

TABLE II
QOE PREDICTION PERFORMANCE OF ST-FL APPROACH.

Synthetic User	Metric	ACR				
		1	2	3	4	5
0	Mean Accuracy	0.77				
	Precision	0.79	0.70	0.73	0.70	0.91
	Recall	0.89	0.63	0.66	0.85	0.80
	F1-Score	0.84	0.66	0.70	0.77	0.85
1	Mean Accuracy	0.76				
	Precision	0.75	0.66	0.73	0.76	0.90
	Recall	0.86	0.59	0.66	0.81	0.89
	F1-Score	0.80	0.63	0.69	0.78	0.89
2	Mean Accuracy	0.77				
	Precision	0.72	0.68	0.78	0.76	0.89
	Recall	0.84	0.61	0.68	0.80	0.90
	F1-Score	0.79	0.60	0.73	0.79	0.90
3	Mean Accuracy	0.76				
	Precision	0.76	0.65	0.79	0.72	0.89
	Recall	0.87	0.63	0.69	0.77	0.85
	F1-Score	0.81	0.64	0.74	0.74	0.87
4	Mean Accuracy	0.74				
	Precision	0.68	0.73	0.70	0.68	0.92
	Recall	0.81	0.60	0.63	0.82	0.83
	F1-Score	0.79	0.60	0.63	0.75	0.87
5	Mean Accuracy	0.75				
	Precision	0.73	0.61	0.74	0.79	0.86
	Recall	0.83	0.60	0.62	0.77	0.93
	F1-Score	0.78	0.60	0.68	0.78	0.89

TABLE III
QOE PREDICTION PERFORMANCE OF THE CB-FL APPROACH.

Synthetic User	Metric	ACR				
		1	2	3	4	5
0	Mean Accuracy	0.79				
	Precision	0.74	0.69	0.80	0.82	0.88
	Recall	0.82	0.62	0.81	0.79	0.89
	F1-Score	0.78	0.65	0.81	0.80	0.88
1	Mean Accuracy	0.79				
	Precision	0.76	0.65	0.76	0.85	0.91
	Recall	0.79	0.60	0.81	0.81	0.94
	F1-Score	0.77	0.62	0.79	0.82	0.93
2	Mean Accuracy	0.87				
	Precision	0.93	0.84	0.78	0.80	0.93
	Recall	0.89	0.87	0.84	0.75	0.93
	F1-Score	0.91	0.85	0.80	0.78	0.93
3	Mean Accuracy	0.87				
	Precision	0.91	0.90	0.75	0.81	0.93
	Recall	0.98	0.84	0.88	0.73	0.89
	F1-Score	0.94	0.87	0.81	0.77	0.91
4	Mean Accuracy	0.76				
	Precision	0.70	0.65	0.81	0.80	0.86
	Recall	0.80	0.60	0.77	0.73	0.93
	F1-Score	0.75	0.62	0.79	0.77	0.89
5	Mean Accuracy	0.81				
	Precision	0.85	0.75	0.67	0.77	0.97
	Recall	0.96	0.61	0.84	0.71	0.86
	F1-Score	0.90	0.67	0.74	0.74	0.91

be seen as this approach achieves greater prediction results for the extreme ACR scores (about 0.7 for 1 and greater than 0.8 for 4 and 5) than for the middle scores (between 0.5 and 0.6 for 2 and 3). This could be due to the imbalanced distribution of the rating scores, i.e., users provided a greater number of extreme scores than middle scores. With regard to the ST-FL approach, a mean accuracy of 0.76 was achieved, computed among the mean accuracy achieved for the 6 synthetic users. Although this result is comparable with that achieved by the Centr-AI approach, the precision, recall and F1-score results highlight that with the ST-FL approach the classification is more balanced than that achieved by the Centr-AI approach; also, the achieved classification values for the ACR scores are greater and always over 0.60. These results may be motivated by the prediction function applied by the FL on the NN weights, which allows to alleviate the dataset imbalance issue. Thus, we can state that, besides preserving the users' privacy, the ST-FL approach also permits to achieve greater QoE prediction precision than the Centr-AI approach.

Finally, we discuss the performance of the proposed CB-FL approach by recalling that two different NNs were trained for this approach, one with the data from synthetic users 0, 1 and 4, and another one with the data from synthetic users 2, 3 and 5. With this approach, the achieved mean accuracy (0.815) is greater than that obtained by the other two approaches; this is particularly true for synthetic users 2, 3 and 5, which achieved better results than synthetic users 0, 1 and 4. This result may be likely due to a better distribution of the ACR scores included in the cluster 26-32. By considering the classification of the single ACR scores, the extreme values (1 and 3) achieved the greatest performance for synthetic users 2, 3 and 5, whereas the score 2 achieved the lowest performance for synthetic users 0, 1 and 4. However, in general for the single ACR classes, the CB-FL approach obtained greater performance than the ST-FL and Centr-AI approaches. Therefore, the results highlight that the CB-FL approach can provide better and more accurate results, also providing predictions which are not biased towards classes with a higher number of samples. Moreover, the synthetic users 4 and 5, which did not contribute to the training stage, explain how the CB-FL approach can generalise better, producing better outcomes for each metric used to evaluate the ACR score prediction.

V. Conclusion

This paper proposes a CB-FL approach applied to QoE modelling. We compare Centr-AI with the ST-FL and the CB-FL approaches to demonstrate that ST-FL can achieve better results than Centr-AI, and our proposed CB-FL can outperform the yet accurate ST-FL results. Moreover, CB-FL improved the inter-classes biased classification problems thanks to its clustering approach, reaching more accurate results for each ACR score prediction, demonstrating its applicability for the QoE modelling task.

The evaluation of the proposed QoE modelling approaches is performed within a single dataset because of the absence of publicly available datasets with the characteristics that this test requires. Thus, tests on CB-FL on other datasets are planned for future works to improve the results and obtain a trustworthy validation of this work.

References

[1] A. Ahmad, A. B. Mansoor, A. A. Barakabitze, A. Hines, L. Atzori, and R. Walshe, "Supervised-learning-Based QoE Prediction of Video Streaming in Future Networks: A Tutorial with Comparative Study," *IEEE Communications Magazine*, vol. 59, no. 11, pp. 88–94, 2021.

[2] Y. Gao, X. Wei, and L. Zhou, "Personalized QoE Improvement for Networking Video Service," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2311–2323, 2020.

[3] P. K. et al., "Advances and Open Problems in Federated Learning," *Found. and Trends in Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[4] S. Ickin, K. Vandikas, and M. Fiedler, "Privacy Preserving QoE Modeling Using Collaborative Learning," in *Proc. of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks*. ACM, 2019, p. 13–18.

[5] L. Amour, S. Sami, S. Hoceini, and A. Mellouk, "Building a Large Dataset for Model-Based QoE Prediction in the Mobile Environment," in *Proc. of the 18th ACM Int. Conf. on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 2015, p. 313–317.

[6] T. Begluk, J. B. Husić, and S. Baraković, "Machine learning-based QoE prediction for video streaming over LTE network," in *2018 17th Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, 2018, pp. 1–5.

[7] T. N. Duc, C. T. Minh, T. P. Xuan, and E. Kamioka, "Convolutional Neural Networks for Continuous QoE Prediction in Video Streaming Services," *IEEE Access*, vol. 8, pp. 116 268–116 278, 2020.

[8] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction," *IEEE Trans. on Multimedia*, vol. 22, no. 12, pp. 3210–3223, 2020.

[9] Y. Wang, P. Li, L. Jiao, Z. Su, N. Cheng, X. Shen, and P. Zhang, "A Data-Driven Architecture for Personalized QoE Management in 5G Wireless Networks," *IEEE Wireless Comm.*, vol. 24, pp. 102–110, 02 2017.

[10] D. Ghadiyaram, J. Pan, and A. C. Bovik, "Learning a Continuous-Time Streaming Video QoE Model," *IEEE Tran. on Image Processing*, vol. 27, no. 5, pp. 2257–2271, 2018.

[11] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous Prediction of Streaming Video QoE Using Dynamic Networks," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1083–1087, 2017.

[12] W. Shi, Y. Sun, and J. Pan, "Continuous Prediction for Quality of Experience in Wireless Video Streaming," *IEEE Access*, vol. 7, pp. 70 343–70 354, 2019.

[13] S. Khanal, K. Thar, and E.-N. Huh, "Route-Based Proactive Content Caching Using Self-Attention in Hierarchical Federated Learning," *IEEE Access*, vol. 10, pp. 29 514–29 527, 2022.

[14] M. Isaksson and K. Norrman, "Secure Federated Learning in 5G Mobile Networks," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[15] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.