APPENDIX

I. HYPERPARAMETER OPTIMISATION

In this section, we detail the hyperparameter optimisation strategies used in our FL experiments. An overview of how we optimized the FL models is reported in Figure 1. Compared to traditional centralised learning frameworks, FL introduces additional challenges, as hyperparameter tuning must consider both standard (canonical) hyperparameters and those specific to distributed learning environments. In centralised learning, hyperparameter optimisation benefits from unrestricted access to the entire dataset, allowing for direct minimisation of a global objective function. In contrast, federated settings require optimisation across a decentralised network of clients, each holding private data with distinct and often non-identically distributed characteristics. In centralised learning, the core hyperparameter optimisation objective is formulated as:

$$\underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathcal{L}(\theta) \tag{1}$$

where $\mathcal{L}(\theta)$ denotes the global training objective.

Depending on the federated setting, cross-silo or cross-device, different optimisation strategies are adopted.

a: Cross-silo

In the cross-silo setting, each client locally evaluates a validation loss, denoted as $\mathcal{L}_i(\theta)$. The global loss is then defined as the average of these local evaluations across the federation:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i(\theta) \tag{2}$$

Here, N denotes the total number of clients. Each $\mathcal{L}_i(\theta)$ is computed on a client-specific validation set, separate from the client's training data. This is possible because in this setting, clients have enough data to split them into the canonical train, validation, and test sets. In our experiments, we adopt the weighted F1-score as the evaluation metric, as it better captures performance under class imbalance and client heterogeneity. Therefore, at the end of the hyperparameter tuning, we select the hyperparameters that maximise the F1-score computed by averaging the F1-scores of each client on its validation set. Then we retrain the federated model on the joint set of training and validation sets of nodes.

Our implementation maintains some consistent architectural choices across the federation: employing ReLU activation functions throughout the network, with a LogSoftmax activation in the final layer, chosen for its numerical stability. The Adam optimiser and categorical cross-entropy loss function provide robust optimisation characteristics across the heterogeneous client data distributions. The resulting hyperparameter configuration

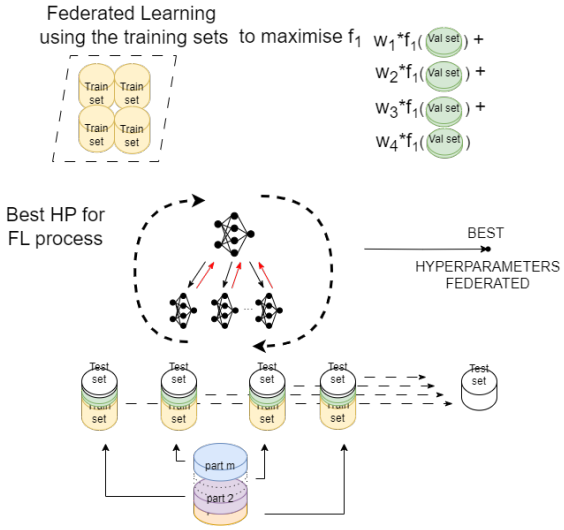

FIGURE 1. The hyperparameter search in a federated context, within the cross-silo scenario. The experimental setting partitions the original dataset into multiple subsets, each assigned to a client. In the cross-silo scenario, each client maintains its own local train, validation, and test sets. In contrast, in the cross-device scenario, clients are split such that some are used for training, others for validation, and the rest for testing.

emerges from a combination of local computation, global aggregation, ultimately yielding models that perform robustly across the federation.

b: Cross-device

In the cross-device scenario, the optimisation process differs significantly. The clients themselves are divided into three sets: training, validation, and test. The training clients train the model that is then validated by the validation set of clients. The hyperparameters are then selected based on the performance on the validation set of clients, aggregating their F1-scores and choosing the model that maximizes this score. Then we retrain the federated model on the union of the training and validation data of the selected clients.

A. FEDERATED HYPERPARAMETER OPTIMISATION STRATEGY

To efficiently navigate the complexity of the distributed optimisation landscape, we adopt a Bayesian optimisation framework. This approach is particularly well-suited to FL scenarios, where each hyperparameter evaluation incurs significant overhead due to the need for coordination among multiple clients. As a result, reducing the number of expensive evaluations becomes critical.

We implement this optimisation process using the Weights & Biases (WandB) platform [**wandb**], enabling the systematic tuning of an extended set of hyperparameters.

In our experiments, the number of local epochs controls how extensively each client trains on its

local dataset before participating in model synchronisation. To stabilise the training process and limit variability, we fix the number of FL communication rounds at twenty.

We tailored the hyperparameter search to the specific characteristics of each federated setting. Since in the cross-silo scenario the clients typically have more stable and resource-rich environments that can support more complex models, we explored a wider range of parameters: number of layers, hidden units, batch size, learning rate, local epochs, and dropout.

In contrast, the cross-device setting focused on batch size, learning rate, local epochs, and optimiser (Adam was the best one across all experiments but one), reflecting the constraints of large-scale, resource-limited client populations where lightweight models and simpler tuning are more practical.

B. BEST HYPERPARAMETERS

We define deep learning models using PyTorch [**NEURIPS2019_9015**], with all neural networks employed being shallow models.

Dataset name	# clients	Number of Layers		
		IID	non-IID $\alpha = 5.0$	non-IID $\alpha = 0.99$
Dry bean	12	2	3	2
	16	2	1	1
Adult	12	3	3	3
	16	2	2	2
Diamond	12	3	1	1
	16	3	2	2

TABLE 1. Best Hyperparameters of layers for clients, distribution and dataset variations for the experiments

Dataset name	# clients	Hidden units		
		IID	non-IID (Medium)	non-IID (High)
Dry bean	12	128	56	90
	16	33	23	109
Adult	12	19	81	45
	16	84	71	58
Diamond	12	75	87	50
	16	123	98	86

TABLE 2. Best Hyperparameters of hidden units for clients, distribution and dataset variations for the experiments

Dataset name	# clients	Batch size		
		IID	non-IID (Medium)	non-IID (High)
Cross-Silo				
Dry bean	12	64	80	128
	16	128	4	80
Adult	12	16	80	8
	16	64	128	128
Diamond	12	64	64	64
	16	32	32	80
Cross-Device				
Dutch	50	82	241	37
	70	181	475	412
	150	316	270	145
ACSI	51		1525	

TABLE 3. Best Hyperparameters of batch size for clients, distribution and dataset variations for the experiments

Dataset name	# clients	Local epochs		
		IID	non-IID (Medium)	non-IID (High)
Cross-Silo				
Dry bean	12	1	2	4
	16	5	4	4
Adult	12	1	2	5
	16	2	4	5
Diamond	12	5	3	5
	16	5	4	5
Cross-Device				
Dutch	50	5	4	5
	70	6	9	8
	150	8	9	8
ACSI	51		5	

TABLE 4. Best Hyperparameters of local epochs for clients, distribution and dataset variations for the experiments

Table 7 presents the best scores obtained with the optimised hyperparameters. For the Dry bean dataset across different client numbers, the F1-score remains relatively stable, indicating the robustness of the optimised model. There are slight variations in the score, with non-IID (Medium) showing slightly better performance. Conversely, for the Adult dataset, the F1-score fluctuates more, especially with non-IID (High) distributions. Interestingly, non-IID (High) occasionally outperforms IID, hinting at possible benefits of heterogeneous data. For the Diamond dataset, the weighted F1-score shows more variation, ranging from 0.755 to 0.607. There is a noticeable drop in performance as the skew level and the number of clients increase. The model's performance degrades with more clients, particularly under skewed distributions, due to the increased heterogeneity of the data.

Regarding the hyperparameter related to the

Dataset name	# clients	Learning rate		
		IID	non-IID (Medium)	non-IID (High)
Cross-Silo				
Dry bean	12	0.127	0.043	0.050
	16	0.338	0.135	0.122
Adult	12	0.165	0.005	0.257
	16	0.250	0.372	0.463
Diamond	12	0.085	0.235	0.112
	16	0.141	0.100	0.203
Cross-Device				
Dutch	50	0.045	0.022	0.010
	70	0.019	0.093	0.049
	150	0.063	0.001	0.060
ACSI	51		0.018	

TABLE 5. Best Hyperparameters of learning rate for clients, distribution and dataset variations for the experiments

Dataset name	# clients	Dropout		
		IID	non-IID (Medium)	non-IID (High)
Dry bean	12	0.180	0.093	0.125
	16	0.035	0.175	0.049
Adult	12	0.006	0.039	0.004
	16	0.019	0.082	0.070
Diamond	12	0.087	0.029	0.044
	16	0.023	0.066	0.014

TABLE 6. Best Hyperparameters of dropout for clients, distribution and dataset variations for the experiments

number of layers, as reported in Table 1, there is no consistent best number. For Dry bean, however, there is a trend in skewed distributions, which often require more layers, reflecting the increased complexity in learning from non-homogeneous data. The Adult dataset shows more consistency with a higher number of layers. The Diamond dataset generally prefers three layers across most conditions, with a reduction in non-IID (Medium), indicating a potential need for deeper networks to capture the nuances of the data. The number of hidden units reported in Table 2 varies widely for the Dry bean dataset, indicating sensitivity to both client number and data distribution. The Adult dataset exhibits more variability, possibly due to the increased complexity of the data. Similarly, the Diamond dataset shows a wide range of hidden units required, with higher numbers for non-IID (High) in some cases, suggesting a need to capture more complex patterns.

The hyperparameter related to batch size is reported in Table 3. For Dry bean, larger batch

Dataset name	# clients	Learning rate		
		IID	non-IID (Medium)	non-IID (High)
Dutch	50	adam	adam	adam
	70	adam	adam	adam
	150	adam	adam	sdg
ACSI	51	adam	adam	adam

TABLE 7. Best Hyperparameters regarding the optimiser across the distribution and dataset variations for the experiments in the setting with Cross-Device.

sizes are generally preferred, with some exceptions for smaller client numbers. In the Adult dataset, batch sizes are mostly large with some variability; skewed distributions commonly require larger batches, potentially stabilising gradients during training. The Diamond dataset prefers smaller batch sizes due to the high number of classes and better gradient estimation from diverse data. In the FL process, the number of local epochs is a crucial hyperparameter, defining how many epochs are performed locally by the participant before sending the updated model to the server. The best number of local epochs is reported in Table 4. For Dry bean, skewed distributions often require more local epochs, indicating slower convergence. Similarly, for Adult, non-IID (High) consistently requires more epochs. The Diamond dataset also shows a preference for more epochs, especially in skewed distributions, suggesting a need for more local updates to achieve stable convergence.

Overall, learning rate (reported in Table 5) and dropout trends (reported in Table 6) indicate that skewed distribution settings benefit from conservative adjustments, such as lower learning rates and dropouts. Skewed data distributions are more complex and require more careful tuning of hyperparameters to achieve good performance. Smaller batch sizes are preferred in skewed distribution settings across all datasets.

Skewed distribution settings generally increase the need for more epochs and smaller batch sizes, indicating that skewed data distributions require more careful and prolonged training to achieve stable and good performance. Higher complexity (more layers) is often needed in skewed distribution settings, suggesting the models need to capture more nuanced data patterns. The Dry bean dataset shows robustness across distributions, with stable performance and fewer drastic changes in hyperparameters. In contrast, the Diamond dataset shows more sensitivity to distributional changes, with notable performance drops and significant hyperparameter adjustments.

Finally, skewed data distributions typically demand less simple training strategies, including more

epochs, smaller batch sizes (see Table 7), and careful tuning of learning rates and dropout rates. These adjustments help maintain model performance despite the increased heterogeneity in data. Understanding these trends can inform better model design and training strategies for diverse real-world datasets.

C. HARDWARE SPECIFICATIONS

The experiments in the cross-silo setting are conducted on a server equipped with an Intel Xeon Gold 6342 processor with 96 cores and 2 Quadro RTX A6000 GPUs. The experiments of the cross-device setting are performed on a server equipped with Dual AMD Rome 7742, 128 cores and 8 A100 GPUs.

II. RESULTS WITH OTHER DATASETS

Table 8 reports the mean Faithfulness and standard deviations for all aggregation strategies, datasets, and distributions in both cross-silo and cross-device settings.

A. CROSS-SILO

In this appendix, we present comprehensive experimental results for our cross-silo FL setup. The experiments were conducted using the Flower [**flower**] framework, with full client participation in each training round. We evaluated our approach across three distinct datasets, each partitioned using multiple strategies to simulate different real-world scenarios.

a: Diamond

The `Diamond` involves classifying diamonds into different quality categories, where each class represents a different quality cut. The dataset contains 10 features, including carat (the weight), clarity, depth, table (width of the top of the diamond relative to the widest point), price, length, width, depth, and colour. Figure 2 presents the results of the different aggregation strategies for Shap explanations in the setting with 12 clients. The figure is composed of three main plots, each corresponding to one of the analysed distributions: (i) an IID setting; (ii) $\alpha = 5.0$ moderately imbalanced, and (iii) $\alpha = 0.99$ highly imbalanced setting. For each scenario, the figure is a Critical Difference plot, in the following referred to as a CD plot. The CD plot is a ranking procedure that we exploit to visualise the statistical differences among the aggregation strategies considered.

In the IID and highly skewed ($\alpha = 0.99$) settings, E_{Q1} ranks consistently first, achieving a mean rank of 1.5 in IID. Notably, E_{faith} and E_{Mix} are not statistically significantly different from each other in these two distributions, both forming part of the top-performing group. In the moderately skewed setting

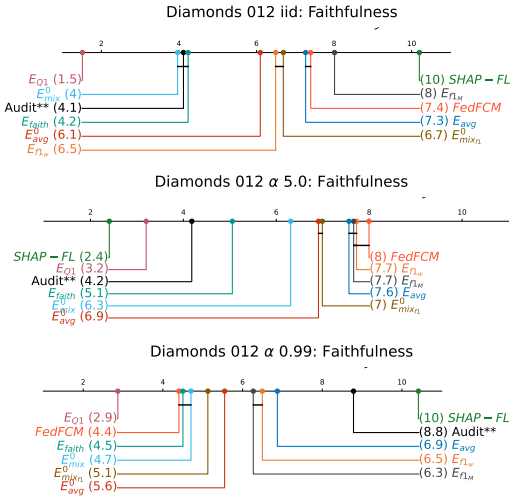

FIGURE 2. Faithfulness CD diagrams of the aggregation strategies, the Audit and the FedFCM on the `Diamond` dataset, in the setting with 12 clients.

($\alpha = 5.0$), the ranking shifts: while E_{Q1} remains competitive, E_{faith} achieves a lower mean rank of 5.1, positioning it in the fourth statistical group, a notable difference from the top-performing methods. This variability suggests that E_{faith}'s performance is more sensitive to moderate imbalance than extreme skewness. The Audit demonstrates extreme sensitivity to distribution characteristics. In IID and $\alpha = 5.0$ settings, it performs well and is part of the top statistical group. However, in the highly skewed case ($\alpha = 0.99$), it deteriorates dramatically to a mean rank of 8.8, while the SHAP-FL competitor shows the poorest performance (rank 10), highlighting that centralized baselines struggle under extreme data imbalance. The competitors FedFCM and our aggregations generally outperform simple averaging approaches based on F1 scores ($E_{F1_{macro}}$, E_{F1}). This indicates that leveraging heterogeneous client contributions, whether through federated aggregation or competitor baselines, provides an advantage over uniform weighting.

All three outperformed averaging approaches based on the F1 score ($E_{F1_{macro}}$,E_{F1}).

In Figure 3, 4 and 5 we present the explanations by class for the `Diamond` dataset, comparing the explanation of the Audit test and the E_{Q1} aggregation. There are the mean explanation values for the samples predicted by class. The blue bars are the mean explanation values of the Audit test by class, and in red are the explanation values of the aggregation E_{Q1}.

Looking at the IID distribution results in Figure 3, we observe strong cosine similarities across classes 0, 1, and 4: in Table 9 there are reported means of all the samples in the explanation set, indicating that E_{Q1} effectively captures the directional trends of FI established by the server. Particularly noteworthy are Classes 0 and 1, showing

Cross-Silo

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}	FedFCM	SHAP-FL	Audit
Dry bean	12	α =0.99	-0.30±0.52	-0.30±0.52	-0.30±0.52	-0.29±0.52	-0.28±0.52	-0.27±0.53	-0.28±0.52	**-0.22**±0.52	-0.29±0.51	-0.31±0.48	-0.28±0.46
		α =5.0	-0.24±0.52	-0.24±0.52	-0.24±0.52	-0.24±0.52	-0.23±0.53	-0.23±0.53	-0.23±0.53	**-0.18**±0.53	-0.24±0.52	-0.27±0.53	-0.25±0.55
		iid	-0.26±0.50	-0.26±0.50	-0.26±0.50	-0.26±0.50	-0.26±0.50	-0.26±0.50	-0.26±0.50	**-0.21**±0.51	-0.26±0.49	-0.26±0.49	-0.27±0.51
	16	α =0.99	-0.21±0.39	-0.21±0.39	-0.21±0.39	-0.21±0.39	-0.20±0.39	-0.20±0.39	-0.20±0.39	**-0.17**±0.39	-0.21±0.41	**-0.17**±0.47	-0.21±0.38
		α =5.0	-0.20±0.55	-0.20±0.55	-0.20±0.55	-0.19±0.55	-0.18±0.55	-0.18±0.56	-0.18±0.55	**-0.15**±0.56	-0.20±0.55	-0.22±0.56	-0.20±0.55
		iid	-0.24±0.55	-0.24±0.55	-0.24±0.55	-0.23±0.55	-0.23±0.55	-0.23±0.56	-0.23±0.55	**-0.17**±0.56	-0.25±0.54	-0.28±0.56	-0.24±0.57
Adult	12	α =0.99	0.51±0.53	0.51±0.53	0.51±0.53	0.52±0.52	0.52±0.52	0.53±0.52	0.52±0.52	**0.56**±0.53	0.51±0.52	0.48±0.56	0.50±0.52
		α =5.0	0.07±0.33	0.07±0.33	0.07±0.33	0.07±0.33	0.07±0.33	0.07±0.33	0.07±0.33	0.08±0.33	0.10±0.42	**0.41**±0.68	0.07±0.33
		iid	0.15±0.38	0.15±0.38	0.15±0.38	0.15±0.37	0.15±0.37	0.15±0.38	0.15±0.37	0.17±0.37	0.22±0.51	**0.59**±0.75	0.16±0.37
	16	α =0.99	0.06±0.29	0.06±0.30	0.06±0.29	0.07±0.29	0.08±0.29	0.08±0.29	0.07±0.29	0.10±0.30	0.11±0.34	**0.65**±0.68	0.06±0.28
		α =5.0	0.12±0.36	0.12±0.36	0.12±0.36	0.13±0.35	0.14±0.36	0.14±0.36	0.14±0.36	0.15±0.36	0.15±0.43	**0.45**±0.67	0.10±0.34
		iid	0.18±0.39	0.18±0.39	0.18±0.39	0.18±0.38	0.19±0.39	0.19±0.39	0.19±0.39	0.21±0.39	0.20±0.44	**0.54**±0.73	0.17±0.38
Diamond	12	α =0.99	0.13±0.39	0.13±0.39	0.13±0.39	0.14±0.38	0.15±0.38	0.15±0.38	0.15±0.38	**0.17**±0.38	0.16±0.39	-0.05±0.42	0.09±0.38
		α =5.0	0.01±0.49	0.01±0.49	0.01±0.49	0.01±0.49	0.01±0.49	0.01±0.49	0.01±0.49	0.04±0.49	-0.00±0.50	**0.48**±0.43	0.03±0.47
		iid	0.21±0.37	0.21±0.37	0.21±0.37	0.22±0.36	0.23±0.35	0.23±0.35	0.23±0.35	**0.28**±0.35	0.21±0.37	0.06±0.41	0.23±0.37
	16	α =0.99	-0.13±0.53	-0.13±0.53	-0.13±0.53	-0.13±0.53	-0.12±0.53	-0.12±0.53	-0.12±0.53	**-0.09**±0.53	-0.11±0.54	-0.25±0.57	-0.11±0.54
		α =5.0	0.04±0.48	0.04±0.48	0.04±0.48	0.04±0.48	0.06±0.48	0.06±0.48	0.06±0.48	0.09±0.48	0.03±0.50	**0.35**±0.46	0.04±0.47
		iid	-0.08±0.46	-0.08±0.46	-0.08±0.46	-0.06±0.46	-0.02±0.46	-0.02±0.47	-0.02±0.46	**0.06**±0.46	0.00±0.47	-0.65±0.42	-0.12±0.38

Cross-Device

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}	Audit
Dutch	50	α = 0.99	0.38±0.46	0.38±0.46	0.37±0.46	0.38±0.46	0.38±0.46	0.37±0.46	0.37±0.46	0.38±0.46	0.35±0.54
		α = 5.0	0.33±0.44	0.33±0.44	0.33±0.44	0.33±0.44	0.33±0.44	0.33±0.44	0.33±0.44	0.32±0.45	0.34±0.44
		iid	0.36±0.44	0.36±0.44	0.36±0.44	0.36±0.44	0.36±0.44	0.36±0.44	0.36±0.44	0.36±0.44	0.34±0.46
	70	α = 0.99	0.33±0.42	0.33±0.42	0.33±0.42	0.33±0.42	0.33±0.42	0.33±0.42	0.33±0.42	0.33±0.42	0.34±0.41
		α = 5.0	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.35±0.43
		iid	0.35±0.43	0.35±0.43	0.35±0.43	0.35±0.43	0.35±0.43	0.35±0.43	0.35±0.43	0.34±0.43	0.33±0.43
	150	α = 0.99	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.47
		α = 5.0	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.45	0.34±0.46
		iid	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.46	0.34±0.45
ACSI	51	*Natural*	0.31±0.61	0.31±0.61	0.31±0.61	0.31±0.61	0.31±0.61	0.31±0.61	0.31±0.61	0.29±0.63	0.31±0.53

TABLE 8. **Faithfulness of the explanations and standard deviations for 2000 explained samples per setting. In the Cross-Silo section, the SHAP-FL competitor is included and shows strong performance on the `Adult` dataset, particularly under balanced (iid) and moderately skewed ($\alpha = 5.0$) distributions, where it achieves the highest faithfulness scores. However, on `Dry bean` and `Diamond`, E_{Q1} remains the most robust strategy across imbalance levels. Negative values, observable for `Dry bean` and some `Diamond` configurations, indicate anti-correlation between the Shap feature-importance ranking and the model's masking response: the features ranked as most important by Shap are not the ones whose removal most degrades model performance. This can arise in multiclass tasks with complex decision boundaries (e.g., `Dry bean` has 7 classes) where the masking baseline poorly captures local feature interactions. In the Cross-Device section, the SHAP-FL competitor was not included as it was not evaluated in that federated setting, and all aggregation methods show similar performance with no clear winner. Importantly, E_{Q1} consistently achieves competitive or leading scores across both settings.**

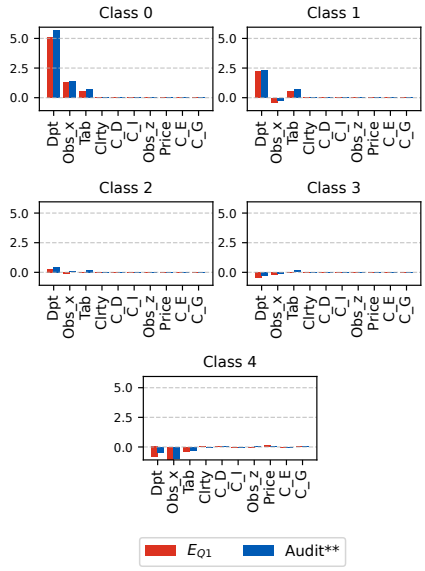

FIGURE 3. **The blue bars are the mean explanation values of the Audit by class, and in red are the explanation values of the aggregation E_{Q1}. They are the global explanation by class: mean of the FI explanations for every predicted class. Here are the explanations for the black box trained in the federation setting with 12 clients in an IID scenario. We show only the first top 10 features in magnitude for readability reasons.**

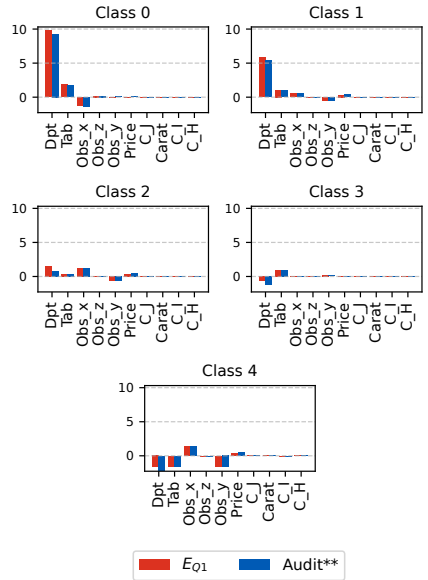

FIGURE 4. **The blue bars are the mean explanation values of the Audit by class, and in red are the explanation values of the aggregation E_{Q1}. They are the global explanation by class: mean of the FI explanations for every predicted class. Here we reported the explanations for the black box trained in the federation setting with 12 clients medium skewed ($\alpha = 5.0$) class distribution.We show only the first top 10 features in magnitude for readability reasons.**

Cross-Silo

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}	FedFCM	SHAP-FL
Dry bean	12	α =0.99	0.97±0.03	0.97±0.03	0.97±0.03	**0.98**±0.03	0.97±0.05	0.97±0.05	0.97±0.05	0.95±0.10	**0.98**±0.02	0.88±0.11
		α =5.0	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.02	**0.99**±0.02	**0.99**±0.02	0.98±0.03	**0.99**±0.02	0.90±0.12
		iid	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	0.99±0.01	0.98±0.02	0.86±0.15
	16	α =0.99	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	0.98±0.04	0.98±0.04	0.98±0.04	0.97±0.04	0.98±0.02	0.86±0.13
		α =5.0	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.01	0.86±0.12
		iid	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	0.99±0.01	0.99±0.02	0.99±0.02	0.99±0.02	0.98±0.02	0.98±0.02	0.83±0.21
Adult	12	α =0.99	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	0.99±0.09	0.99±0.09	0.99±0.09	**1.00**±0.01	**1.00**±0.00	0.20±0.19
		α =5.0	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	0.98±0.07	0.69±0.22
		iid	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	0.95±0.14	0.28±0.19
	16	α =0.99	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	**1.00**±0.00	**1.00**±0.07	0.99±0.07	**1.00**±0.07	**1.00**±0.02	0.98±0.05	0.27±0.19
		α =5.0	**1.00**±0.01	**1.00**±0.02	**1.00**±0.01	**1.00**±0.01	0.99±0.04	0.99±0.04	0.99±0.04	**1.00**±0.02	0.98±0.08	0.70±0.19
		iid	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	**1.00**±0.00	0.99±0.07	0.99±0.07	0.99±0.07	**1.00**±0.01	0.98±0.07	0.56±0.20
Diamond	12	α =0.99	**0.97**±0.07	**0.97**±0.07	**0.97**±0.07	**0.97**±0.08	0.96±0.11	0.96±0.11	0.96±0.11	0.94±0.12	0.88±0.24	0.93±0.12
		α =5.0	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.02	**0.99**±0.02	**0.99**±0.02	**0.99**±0.03	0.97±0.06	0.57±0.15
		iid	0.98±0.03	0.98±0.03	0.98±0.03	0.98±0.04	0.98±0.07	0.98±0.07	0.98±0.07	0.97±0.07	**1.00**±0.01	0.93±0.10
	16	α =0.99	0.96±0.06	**0.97**±0.05	**0.97**±0.05	0.96±0.05	**0.97**±0.05	**0.97**±0.05	**0.97**±0.05	0.96±0.07	**0.97**±0.07	0.87±0.19
		α =5.0	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	0.99±0.04	0.99±0.04	0.99±0.04	0.98±0.04	0.95±0.09	0.81±0.14
		iid	**0.83**±0.22	**0.83**±0.22	**0.83**±0.22	0.82±0.24	0.80±0.26	0.80±0.26	0.80±0.26	0.77±0.29	0.70±0.29	0.41±0.23

Cross-Device

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}
Dutch	50	α =0.99	0.88±0.10	0.88±0.10	**0.89**±0.10	0.88±0.10	0.88±0.10	**0.89**±0.10	**0.89**±0.10	0.88±0.11
		α =5.0	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03	**0.96**±0.03
		iid	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	0.96±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	0.96±0.03
	70	α =0.99	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06	**0.94**±0.06
		α =5.0	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02	**0.98**±0.02
		iid	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03	**0.97**±0.03
	150	α =0.99	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01
		α =5.0	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.01	**1.00**±0.01	**1.00**±0.01	**1.00**±0.00
		iid	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00	**1.00**±0.00
ACSI	51	*Natural*	**0.96**±0.02	**0.96**±0.02	**0.96**±0.02	**0.96**±0.02	**0.96**±0.02	**0.96**±0.02	**0.96**±0.02	**0.96**±0.04

TABLE 9. ***ShapGap*$_{Cos}$ between aggregation strategies and audit baseline across cross-silo and cross-device settings. Values closer to 1.0 indicate better alignment with the server explanation. The SHAP-FL competitor column (cross-silo only) shows relatively low *ShapGap*$_{Cos}$, particularly on `Adult` and `Diamond`, suggesting different explanation patterns compared to federated aggregations.**

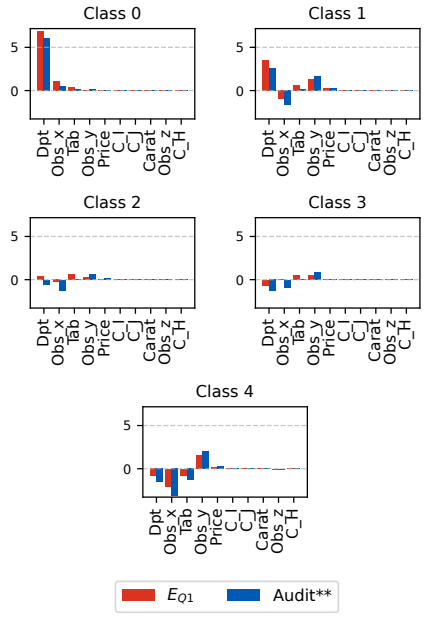

FIGURE 5. **Here are the explanations for the black box trained in the federation setting with 12 clients in a highly skewed scenario. The blue bars are the mean explanation values of the Audit by class, and in red the explanation values of the aggregation E_{Q1}. Classes 0, 1, and 4 have high alignment, but across all classes, there are large absolute gaps. We show only the first top 10 features in magnitude for readability reasons.**

near-perfect alignment.

When examining the $\alpha = 5.0$ setting Figure 4, we notice a shift. The similarities demonstrate even stronger alignment, with some perfect matches for classes 0 and 1, and still very good alignment for the remaining classes. However, as reported in Table 10, the *ShapGap*$_{L2}$ in the medium skewed distribution shows more instability compared to the IID case. This suggests that under moderate imbalance, E_{Q1} maintains or even improves directional accuracy while experiencing greater magnitude variations (due to the nature of the heterogeneous data). The $\alpha = 0.99$ scenario presented in Figure 5 reveals some distinctive patterns: while maintaining high cosine similarities for classes 0, 1, and 4, we observe noticeably larger absolute gaps across all classes. This increase in difference, particularly notable in Classes 2 and 4, suggests that severe imbalance affects the scale of importance values more significantly than their direction.

Looking at this pattern of results across different distributions and classes, we observe a pattern characteristic of the diamond classification problem with varying FI. As `Diamond` involves classifying diamonds into different quality cuts, the varying patterns in classes 2-4 with more distributed and less stable FI represent multiple features that interact to determine the classification. This may explain why these classes show more sensitivity to distribution changes. Class 0 and Class 1's strong reliance

Cross-Silo

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}	FedFCM	SHAP-FL
Dry bean	12	α =0.99	1.02±0.11	1.02±0.11	1.01±0.11	0.97±0.16	1.05±0.26	1.05±0.27	1.03±0.26	1.16±0.53	**0.90**±0.11	2.18±0.14
		α =5.0	**0.41**±0.06	**0.41**±0.06	**0.41**±0.06	0.42±0.06	0.43±0.12	0.43±0.12	0.43±0.12	0.62±0.19	0.55±0.05	1.47±0.08
		iid	**0.26**±0.04	**0.26**±0.04	**0.26**±0.04	0.27±0.05	0.27±0.09	0.27±0.09	0.27±0.09	0.53±0.17	0.74±0.12	1.91±0.22
	16	α =0.99	0.61±0.06	**0.60**±0.06	0.61±0.06	0.61±0.08	0.70±0.33	0.70±0.33	0.70±0.33	1.07±0.30	0.83±0.08	2.36±0.25
		α =5.0	**0.79**±0.07	**0.79**±0.07	0.80±0.07	0.80±0.08	0.83±0.17	0.83±0.17	0.83±0.17	1.07±0.24	0.96±0.11	2.53±0.27
		iid	**0.47**±0.04	**0.47**±0.04	**0.47**±0.04	0.50±0.07	0.50±0.14	0.49±0.14	0.50±0.14	0.85±0.29	0.85±0.12	2.63±0.37
Adult	12	α =0.99	**0.07**±0.03	0.11±0.02	**0.07**±0.03	0.10±0.03	0.09±0.10	0.13±0.11	0.09±0.10	0.36±0.10	0.10±0.05	1.54±0.05
		α =5.0	**0.06**±0.00	**0.06**±0.00	0.07±0.00	0.07±0.01	0.07±0.02	0.08±0.02	0.07±0.02	0.16±0.05	0.53±0.02	1.28±0.03
		iid	**0.05**±0.02	**0.05**±0.02	**0.05**±0.02	**0.05**±0.02	**0.05**±0.02	**0.05**±0.02	**0.05**±0.02	0.06±0.04	0.89±0.02	1.53±0.04
	16	α =0.99	**0.14**±0.02	0.19±0.02	**0.14**±0.02	**0.14**±0.02	**0.14**±0.06	0.15±0.07	**0.14**±0.06	0.21±0.04	0.46±0.05	1.30±0.06
		α =5.0	0.23±0.03	0.23±0.03	**0.22**±0.03	**0.22**±0.02	0.23±0.09	0.23±0.09	**0.22**±0.09	0.24±0.13	0.63±0.02	1.35±0.02
		iid	**0.20**±0.03	**0.20**±0.03	**0.20**±0.03	**0.20**±0.03	0.21±0.13	0.22±0.13	0.22±0.13	0.34±0.16	0.60±0.03	1.59±0.02
Diamond	12	α =0.99	**1.20**±0.29	1.22±0.29	1.22±0.29	1.21±0.29	1.29±0.39	1.30±0.39	1.30±0.38	1.50±0.49	2.28±0.49	1.74±0.34
		α =5.0	0.64±0.11	0.64±0.11	0.64±0.11	**0.63**±0.12	0.65±0.14	0.65±0.14	0.65±0.14	0.76±0.23	1.35±0.28	7.77±0.30
		iid	0.49±0.10	0.49±0.10	0.49±0.10	0.49±0.10	0.50±0.13	0.49±0.13	0.50±0.13	0.49±0.22	**0.22**±0.06	0.93±0.17
	16	α =0.99	0.87±0.12	0.84±0.12	0.85±0.12	0.86±0.12	0.83±0.18	0.82±0.18	0.82±0.18	0.85±0.31	**0.68**±0.21	1.54±0.26
		α =5.0	**0.38**±0.07	**0.38**±0.07	**0.38**±0.07	0.39±0.08	0.49±0.26	0.49±0.26	0.49±0.26	0.82±0.24	1.36±0.16	3.11±0.22
		iid	**0.85**±0.08	**0.85**±0.08	**0.85**±0.08	**0.85**±0.08	0.87±0.10	0.87±0.10	0.87±0.10	0.89±0.13	1.04±0.09	3.61±0.30

Cross-Device

Dataset	Clients	Distribution	E_{Avg}	E_{F1}	$E_{F1_{macro}}$	E_{faith}	E^0_{Avg}	E_{Mix}	$E_{Mix_{F1}}$	E_{Q1}
Dutch	50	α =0.99	0.50±0.23	0.50±0.24	**0.49**±0.24	0.50±0.23	0.50±0.23	**0.49**±0.23	**0.49**±0.24	0.50±0.24
		α =5.0	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10	**0.26**±0.10
		iid	**0.26**±0.12	**0.26**±0.13	**0.26**±0.13	**0.26**±0.12	**0.26**±0.12	**0.26**±0.12	**0.26**±0.13	0.27±0.12
	70	α =0.99	**0.37**±0.08	**0.37**±0.08	**0.37**±0.08	**0.37**±0.08	**0.37**±0.08	**0.37**±0.08	**0.37**±0.08	**0.37**±0.09
		α =5.0	**0.23**±0.02	**0.23**±0.02	**0.23**±0.02	**0.23**±0.02	**0.23**±0.02	**0.23**±0.02	**0.23**±0.02	**0.23**±0.03
		iid	0.25±0.12	0.25±0.12	0.25±0.12	**0.24**±0.12	0.25±0.12	0.25±0.12	0.25±0.12	0.26±0.11
	150	α =0.99	0.16±0.02	0.16±0.02	0.16±0.02	0.16±0.02	0.16±0.02	0.16±0.02	0.16±0.02	**0.15**±0.03
		α =5.0	**0.07**±0.02	**0.07**±0.02	**0.07**±0.02	**0.07**±0.02	**0.07**±0.02	**0.07**±0.02	**0.07**±0.02	0.08±0.03
		iid	0.08±0.02	**0.07**±0.02	**0.07**±0.02	0.08±0.02	0.08±0.02	0.08±0.02	0.08±0.02	0.08±0.02
ACSI	51	*Natural*	**0.28**±0.05	**0.28**±0.05	0.29±0.05	0.29±0.05	**0.28**±0.05	**0.28**±0.05	0.29±0.05	0.29±0.09

TABLE 10. ***ShapGap*$_{L2}$ distance between aggregation strategies and audit baseline across cross-silo and cross-device settings. Lower values indicate better magnitude preservation. The SHAP-FL competitor (cross-silo only) exhibits significantly higher *ShapGap*$_{L2}$ distances, suggesting it produces explanation feature rankings with substantially different magnitude scaling compared to the aggregated federated explanations.**

on depth (total depth percentage) suggests that it is a crucial discriminating factor for the highest quality diamonds, according to the explained black box models. High similarities and low absolute gaps in those classes indicate that E_{Q1} has a robust, well-learned pattern. The varying patterns in classes 2-4, with more distributed and less stable FI, represent multiple features that interact to determine the classification. This would explain why these classes show more sensitivity to distribution changes.

b: Diamond *and* Dry bean *16 client scenario*

With 16 clients, the Diamond rankings largely maintain their 12 client pattern in the IID and highly skewed scenarios, where E_{Q1} remains the top performer. However, a notable shift occurs in the moderately skewed case ($\alpha = 5.0$): while E_{Q1} achieves a competitive rank of 2.5, the SHAP-FL competitor rises to claim superior performance in this intermediate imbalance regime. This finding parallels the Adult dataset behavior, suggesting that moderate imbalance specifically favors competitor-based aggregation strategies across multiple datasets when scaling to 16 clients.

In the 16 client configuration (Figure 9), the scaling effect becomes more pronounced. In the highly skewed setting ($\alpha = 0.99$), the SHAP-FL competitor achieves notably higher faithfulness than in the 12 client case, raising its competitive advantage. Here, FedFCM pairs with E_{Q1} in a second statistical cluster, suggesting that additional client diversity at the 16 client scale provides sufficient heterogeneity for simpler aggregation strategies to reach competitive performance. In the moderately skewed distribution ($\alpha = 5.0$) and IID settings, both SHAP-FL and E_{Q1} rank among the best performers, while FedFCM consistently occupies the second group. This pattern suggests that as federated systems scale, the distinction between sophis-

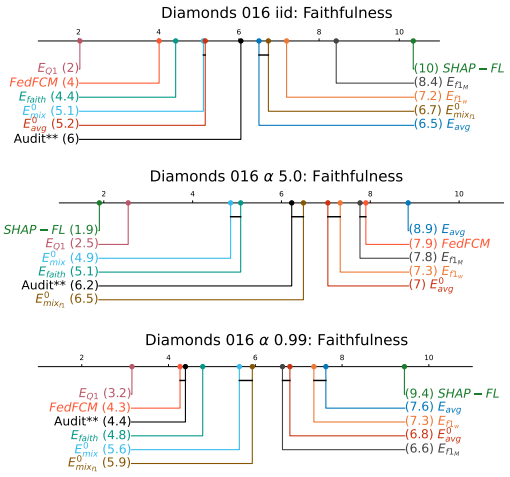

FIGURE 6. **CD diagrams for the Diamond dataset with 16 clients across distributions: (i) IID, (ii) moderately skewed (α = 5.0), and (iii) highly skewed (α = 0.99).**

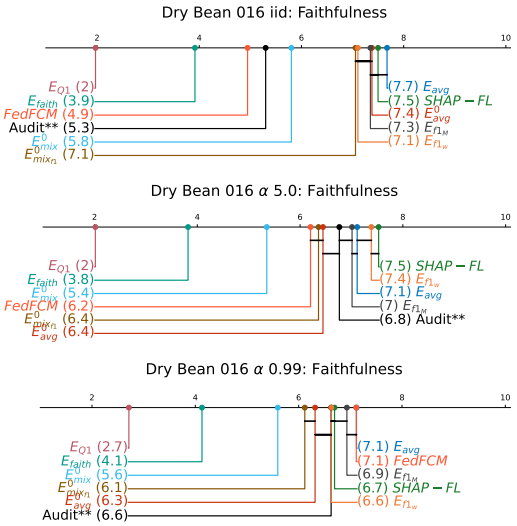

FIGURE 7. CD diagrams for the `Dry bean` dataset with 16 clients across distributions: (i) IID, (ii) moderately skewed ($\alpha = 5.0$), and (iii) highly skewed ($\alpha = 0.99$).

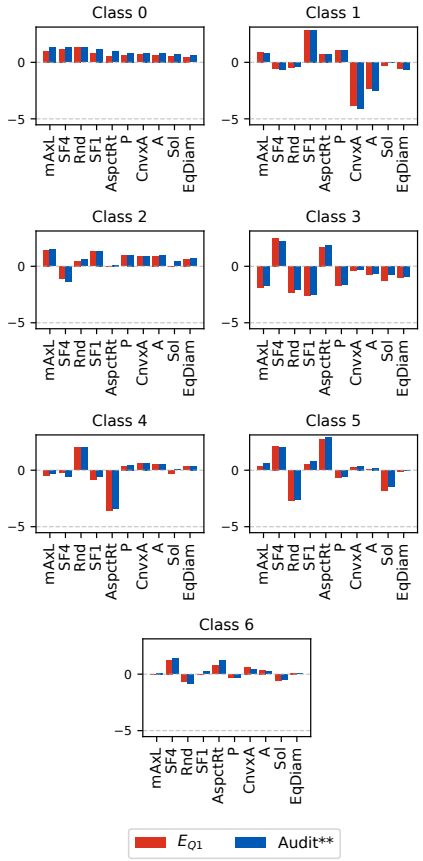

FIGURE 8. Global FI explanations by class for the `Dry bean` dataset with 16 clients in a highly skewed scenario ($\alpha = 0.99$). The blue bars represent Audit, red bars represent E_{Q1}. While the FI magnitudes appear similar, the CD diagram analysis reveals that E_{Q1} achieves a significantly better mean rank through superior consistency across the internal class distributions, demonstrating marked robustness gain over the SHAP-FL competitor under extreme imbalance.

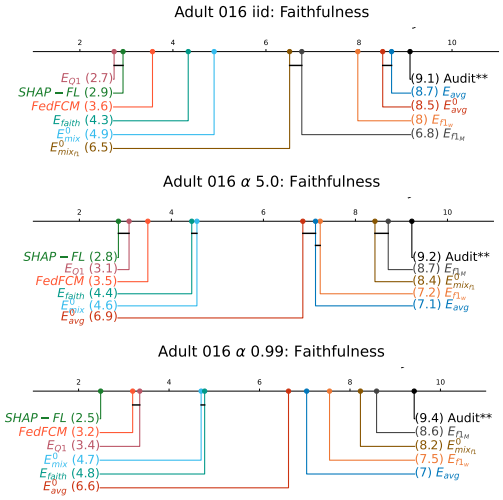

FIGURE 9. CD diagrams for the `Adult` dataset with 16 clients across distributions: (i) IID, (ii) moderately skewed ($\alpha = 5.0$), and (iii) skewed ($\alpha = 0.99$).

ticated and simple aggregation strategies narrows on the `Adult` dataset, a finding with important implications for practitioners choosing aggregation schemes under resource constraints.

c: Adult

`Adult` is a tabular dataset containing demographic information about individuals, such as age, education, and occupation. The aim is to predict whether an individual earns more than $50,000 a year (binary classification). The Figure 10 presents the results of the different aggregation strategies for Shap explanations in the setting with 12 clients. The SHAP-FL competitor demonstrates superior performance on the `Adult` dataset across multiple distributions. In the IID setting, SHAP-FL achieves the best faithfulness with E_{Q1} and FedFCM forming a second statistical group (both with higher mean ranks than SHAP-FL). Specifically, E_{Q1} achieves a mean rank of 3.3 in the IID setting and improves substantially to 1.6 in the highly skewed ($\alpha = 0.99$) setting, where it becomes the top performer. In the moderately skewed distribution ($\alpha = 5.0$), SHAP-FL again ranks best with 2.9, followed by FedFCM (rank 3.5) in the second group, suggesting that this intermediate imbalance may particularly favor the competitor's explanation aggregation strategy. Notably, FedFCM does not outperform our E_{Q1} approach; instead, they operate in different performance tiers depending on distribution. In the highly skewed setting ($\alpha = 0.99$), E_{Q1} achieves clear dominance (rank 1.6), while E_{faith} and E_{Mix} form a second statistical group (ranks 3.4 and 3.6), both providing competitive alternatives. The Audit retreats to the worst-performing group in this extreme imbalance case. This pattern demonstrates that federated aggregation leveraging client heterogeneity provides robustness under distribution shift.

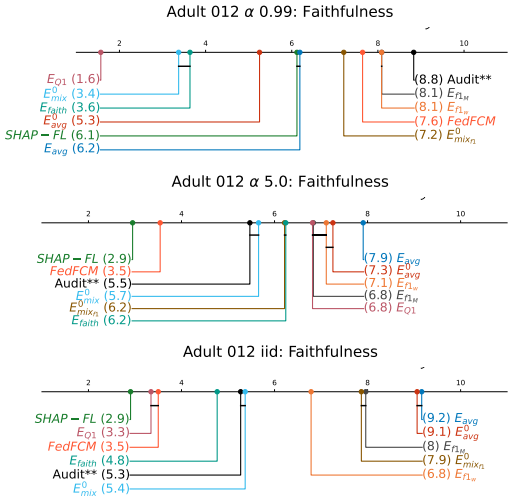

FIGURE 10. Faithfulness values for the different aggregation strategies on the `Adult` dataset, in the setting with 12 clients across the different experimental distributions.

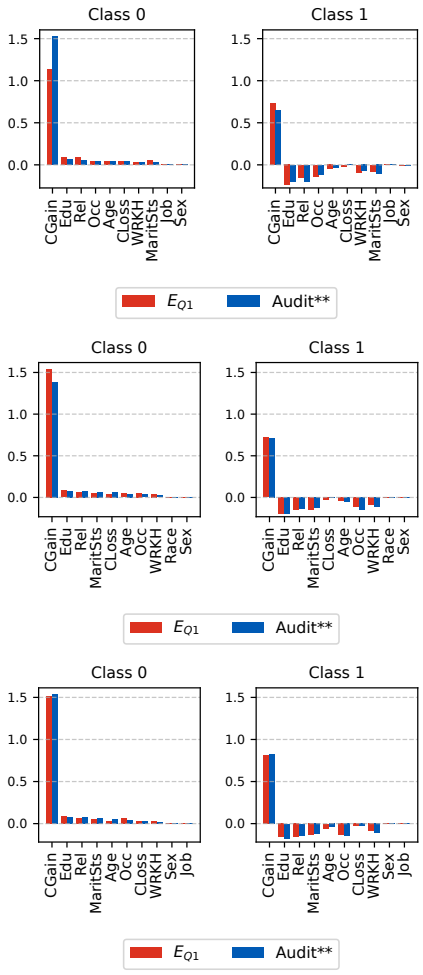

FIGURE 11. Global FI explanations by class for the `Adult` dataset, in the setting with 12 clients. The blue bars are the mean explanation values for the samples predicted by class according to the audit explanation. For both classes, the $ShapGap_{L2}$ and $ShapGap_{Cos}$ are reported. The similarity in terms of $ShapGap_{Cos}$ is very high for both classes in all the distributions. The $ShapGap_{L2}$ grows when the skewness increases: as we move from the most skewed distribution to the IID one, the $ShapGap_{L2}$ decreases for the explanations of both classes. We show only the first top 10 features in magnitude for readability reasons.

By looking at Figure 11, we can see the global explanations by class for the `Adult` dataset. The blue bars are the mean explanation values for the samples predicted by class according to the Audit explanation. The means of $ShapGap_{L2}$ and $ShapGap_{Cos}$ of both the classes are reported in Table 10 and Table 9. The alignment of sign and magnitude is very high for both classes in all the distributions. E_{Q1} captures the directional importance of features that influence the federated model. Instead, looking at the difference metric, we see an expected pattern: the increase in skewness leads to a higher difference in the magnitude of difference of FI explanations: the $ShapGap_{L2}$ grows when the skewness increases. As we move from the most skewed distribution to the IID one, the $ShapGap_{L2}$ decreases, for the explanation of both classes. The feature *capital gain* consistently emerges as the most important feature across all distributions and classes, with the highest magnitude of importance.

B. CROSS-DEVICE

In this setting, we assume to be in a cross-device scenario, dividing clients into two sets, one for the train and one for the test. On each FL round, we sample 30% of the training clients.

a: Income

`ACSI` [**ding2022retiring**] is a tabular dataset built from the American Community Survey (*ACS*) with data coming from all 50 states and Puerto Rico in 2018. The task involves predicting whether the samples in the dataset have a salary higher than $50,000. Since it is split into 51 parts based on the state, we used the natural partition as clients in our experiments. This allows us to validate our method in this setting.

The Critical Difference diagram of the Faithfulness is in Figure 13. The E_{Q1} maintains its primacy among other methods in terms of Faithfulness: this does not surprise us and is in line with our expectations. This means that also in the cross-device setting, with a natural partitioning, the E_{Q1} is the best aggregation strategy. Also for the `ACSI` dataset, the aggregation strategies that take into account the Faithfulness of the clients, E_{faith} and E_{Mix}, are the second and third best-performing strategies. The Audit test achieves the highest Faithfulness values (alongside E_{Q1}), whereas in the other experimental datasets (see Paragraph II-A0c and II-A0a), its performance fluctuates around the mean ranks. This variation suggests that the nature of the data plays a crucial role in determining the relative effectiveness of these methods. In the bar plot of Figure 14 we

FIGURE 12. **Global FI explanations by class for the `Adult` dataset using the SHAP-FL competitor method, in the setting with 12 clients. Comparison with Figure 11 shows that SHAP-FL achieves notably higher faithfulness scores (particularly under moderate and balanced distributions), demonstrating competitive robustness as a federated explanation baseline.**

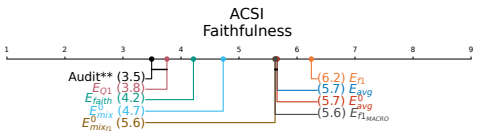

FIGURE 13. **Critical Difference diagram of the Faithfulness of the different aggregations for the `ACSI` dataset.**

present the global FI explanations per class of the E_{Q1} aggregation, red bars, for the `ACSI` dataset, compared with the blue bars of the Audit test. The $ShapGap_{Cos}$ from Table 9 shows moderately strong alignment between E_{Q1} and the Audit explanations for both classes, with a mean value around 0.85 for Class 0 and Class 1. This suggests that our aggregation policy captures the directional importance of features consistently across classes. However, the absolute difference metric lets the differences emerge. While the directional alignment might be similar or slightly better for Class 1, the absolute magnitude of FI is more accurately preserved for

FIGURE 14. **The global explanation by class for the dataset `ACSI`.**

Class 0. This divergence between directional and magnitude accuracy is significant from a methodological perspective. Looking at the FI, we can see that Class 0 exhibits bigger importance values for certain key features, notably `WKHP` (*working hours*) and SEX, while Class 1 shows more importance values with negative contributions. This pattern explains the higher L2 gap for Class 1, as more complex feature interaction patterns could be harder to capture through aggregation. These findings suggest that while E_{Q1} performs capturing the overall explanation structure, its effectiveness varies between classes in ways that might be relevant for practical applications. The trade-off between directional accuracy and magnitude preservation often appears to be class-dependent, which could have implications for how we interpret these FI explanations in practice.

b: Dutch, other cross-device configurations

For completeness, we report below the CD diagrams for the `Dutch` dataset with 150 client configurations. In contrast to the 70 client case discussed in the main paper, increasing the number of clients to 150 significantly changes the ranking of aggregation strategies. With many clients, the advantage of E_{Q1} diminishes, and simple averaging (E_{Avg}) and averaging with filtering (E^0_{Avg}) become competitive, demonstrating that when sufficient client diversity is achieved, advanced weighting schemes provide marginal gains. In highly skewed distributions ($\alpha = 0.99$), aggregation methods that weight by local model performance ($E_{F1_{macro}}$ and $E_{Mix_{F1}}$) emerge as the top performers, indicating that under extreme imbalance, balancing the local goodness of explanations becomes more important than preserving client-weighted rankings.

III. CHOICE FOR MASKING VALUE FOR FAITHFULNESS ESTIMATION

The choice of masking value significantly impacts Faithfulness evaluation, as it determines the base-

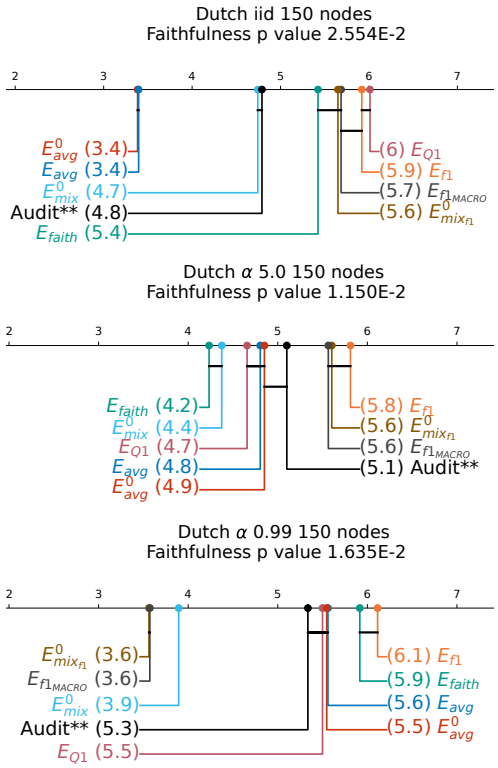

FIGURE 15. **CD diagrams for the `Dutch` dataset with 150 clients across distributions: (i) IID, (ii) moderately skewed ($\alpha = 5.0$), and (iii) highly skewed ($\alpha = 0.99$).**

line against which feature contributions are measured [1]. Literature shows that distributional baselines generally outperform constant values for feature attribution methods [**zheng2024f**]. Candidate Values Tested: We evaluated multiple masking approaches: fixed values (-1, -0.5, 0.001, 0.5, 1.0), local column-wise mean, and local dataset mean. Figure 16 shows Faithfulness distributions across the experiments; The local total average consistently provided the most stable results, offering: (i) distributional consistency avoiding out-of-distribution artifacts, (ii) privacy preservation using only local data, (iii) theoretical alignment with interventional SHAP principles, and (iv) robustness to federated data heterogeneity. Fixed values fell outside natural feature ranges, creating unrealistic inputs. Column-wise means added complexity without proportional benefits. Global statistics would violate federated privacy constraints. We selected the local total average as it balances efficiency, privacy, and methodological soundness while following established best practices for distributional baseline approaches in explainable AI.

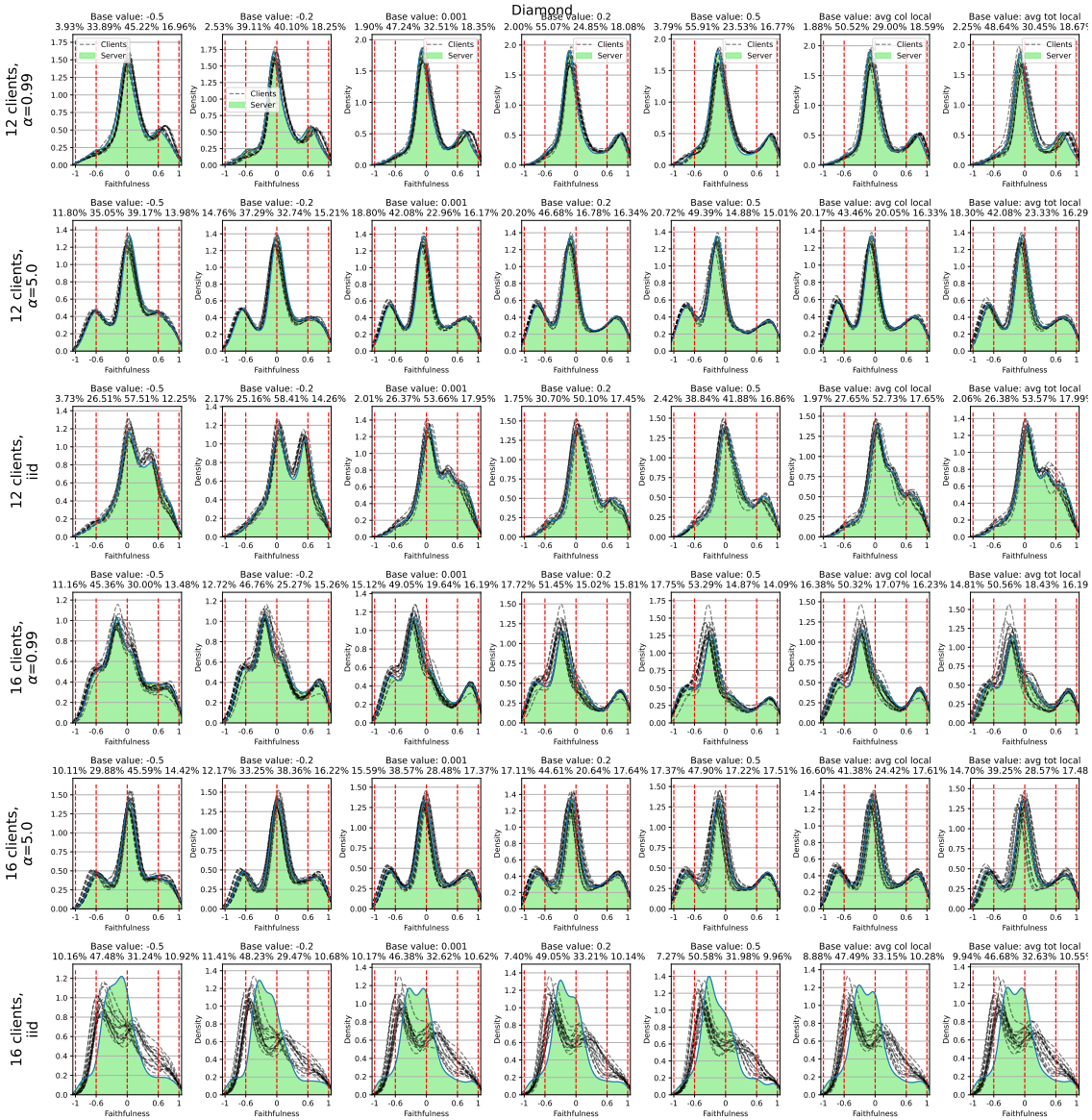

FIGURE 16. **In every subplot, there are the estimated density distributions of the Faithfulness values of the clients (dashed lines). The estimated distribution of the Faithfulness values of the Audit test is plotted with the solid color. Each row displays the values for the setting, composed of the number of clients and the distribution. In each column, there is a different base value choice. Only the ablation for the `Diamond` is reported here, but the other datasets presented comparable results.**