UNIVERSITÀ DI PISA

Consistent Post-Hoc Explainability in Federated Learning through Federated Fuzzy Clustering

Pietro Ducange, Francesco Marcelloni, Alessandro Renda, Fabrizio Ruffini

University of Pisa, Dept. of Information Engineering, Pisa, Italy

International Conference on Fuzzy Systems (FUZZ-IEEE) IEEE WCCI, June 30-July 5, 2024

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CROSSLAB Innovation for industry 4.0

FoReLab Rethinking the future

IEEE Computational Intelligence Society

WCCI 2024 Yokohama Japan

Outline

- **Introduction and motivation**
  - The pursuit of trustworthiness
  - **Fed-XAI**: Federated Learning of eXplainable AI models
  - Challenges of adopting **SHAP** as post-hoc method in the Federated Learning setting
- **Contribution**
  - **Federated SHAP:** consistent explainability through Federated Fuzzy Clustering
- **Experimental analysis**
  - **Comparison** with alternative approaches and baselines

The pursuit of *trustworthiness*

Human agency and oversight

Accountability

Technical robustness and safety

Societal and environmental wellbeing

Seven requirements toward Trustworthy AI

Diversity non-discrimination and fairness

Privacy and data governance

Transparency

Need to collect data to train accurate ML models clashes with need to preserve privacy of data owners

"AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.""

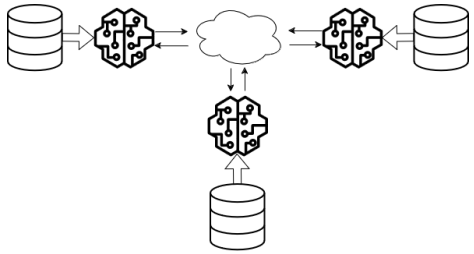
Fed-XAI
Federated Learning of **eXplainable AI models**

Fed-XAI background

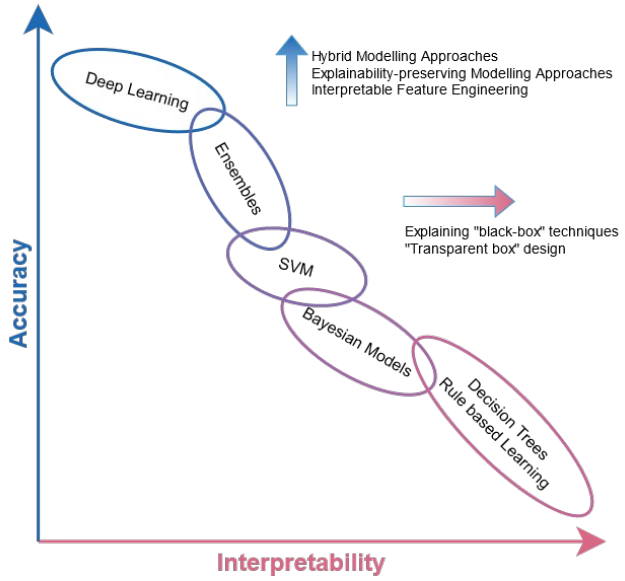
Federated Learning

Federated Averaging (iterates over following steps):

1. Server sends global model to clients
2. Each client updates the model using local data
3. Each client sends the model back to the server
4. Server takes the average of the locally computed updates, weighted according to the number of samples



eXplainable AI



Note

Federated Averaging immediately suitable for *Neural Networks* generally deemed as "opaque" or "black boxes"

Motivation

Note
Federated Averaging immediately suitable for *Neural Networks* generally deemed as "opaque" or "black boxes"

→ How to achieve the **Fed-XAI goal,** i.e., **explainability** in **FL**?

State of art

- Ad-hoc strategies for **FL of inherently interpretable** models → 😓 Possibly, less accurate than *black boxes* for certain tasks
  - e.g., **TSK Fuzzy Rule-Based Systems**
    - Wilbik et al., ***Towards a Federated Fuzzy Learning System*** (2021)

- **Post-hoc explainability techniques** in the FL setting → 😓 Hard to ensure:
  - **privacy** preservation
  - **accurate** & **consistent** explanations
  - e.g., through **SHAP** (SHapley Additive exPlanations)

Our proposal

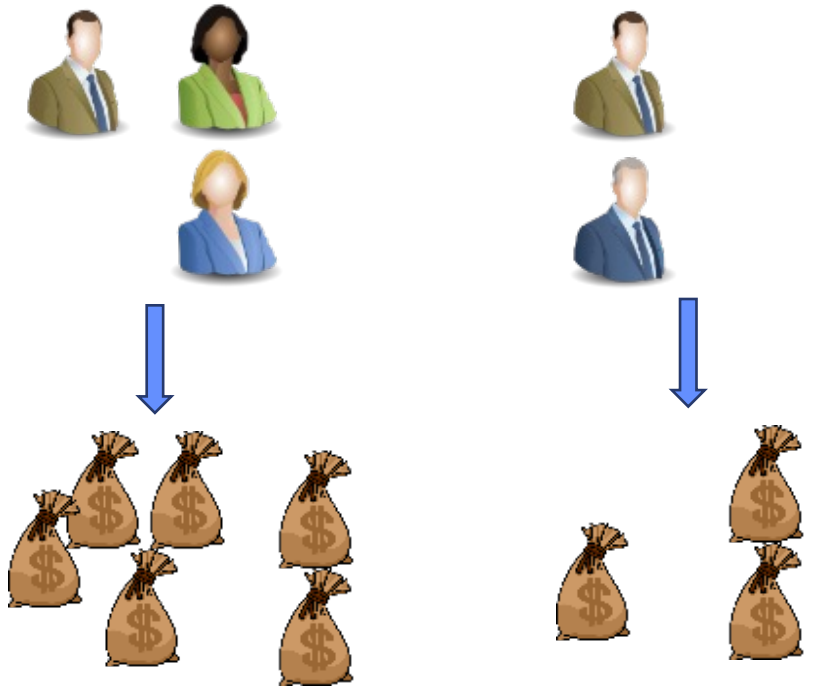
- **FederatedSHAP**: Consistent post-hoc explainability in FL through federated fuzzy clustering

SHAP: Background

Post-hoc method for estimating the **Shapley values** Lloyd Shapley, Nobel Memorial Prize in Economics in 2012

- Intuition: the company-revenue example as a realization of **cooperative game theory**

How to **allocate the payout** among employees?



Lundberg et al. ***A unified approach to interpreting model predictions*** (2017)

SHAP: Background

- From the company-revenue example to **XAI**
  - Replace employees with features
  - Replace profit with model prediction
  - Shapley values quantify the impact of each feature on model prediction

Shapley Values in XAI

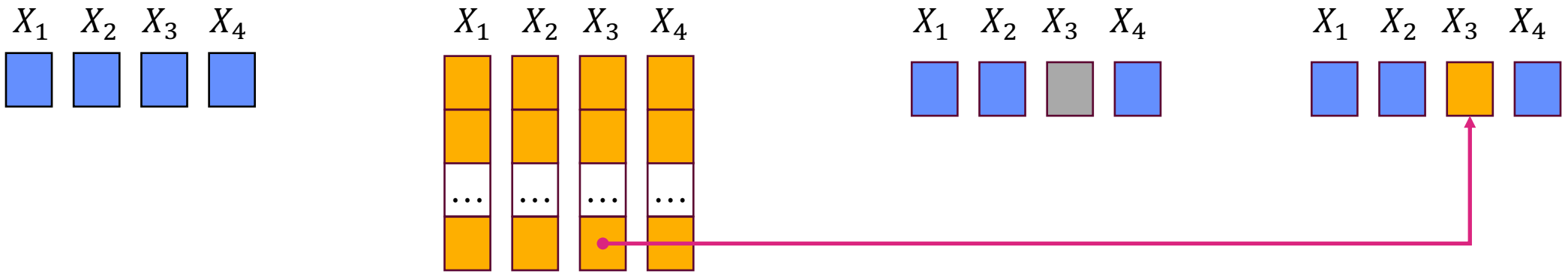
$$\hat{y}_i = f(x_i) = \phi_0 + \sum_{j=1}^{F} \phi_j$$

- f Predictive model
- x_i Generic F-dimensional input instance
- ϕ_0 Average of the predictions from a *background dataset*
- ϕ_i Shapley values

- **Local,** i.e., explains individual predictions
- **KernelShap** variant: linear regression-based approximation
  - More **efficient** than naive calculation
  - **Model-agnostic,** suited for both classification and regression tasks

Challenge of adopting SHAP in the FL setting

- Estimation of Shapley values for the explanations for x_i involves testing *coalitions* of features by *perturbing* x_i
- A **background dataset (BG)** is exploited for perturbing x_i
  - Replace features excluded from a coalition with those of instances randomly sampled from BG
  - The BG should coincide with the set of data used for learning the f model (i.e., the **training set**)
  - It is a common practice to reduce the numerosity of the BG (e.g., through sampling)



Challenge of adopting SHAP in the FL setting

Challenges

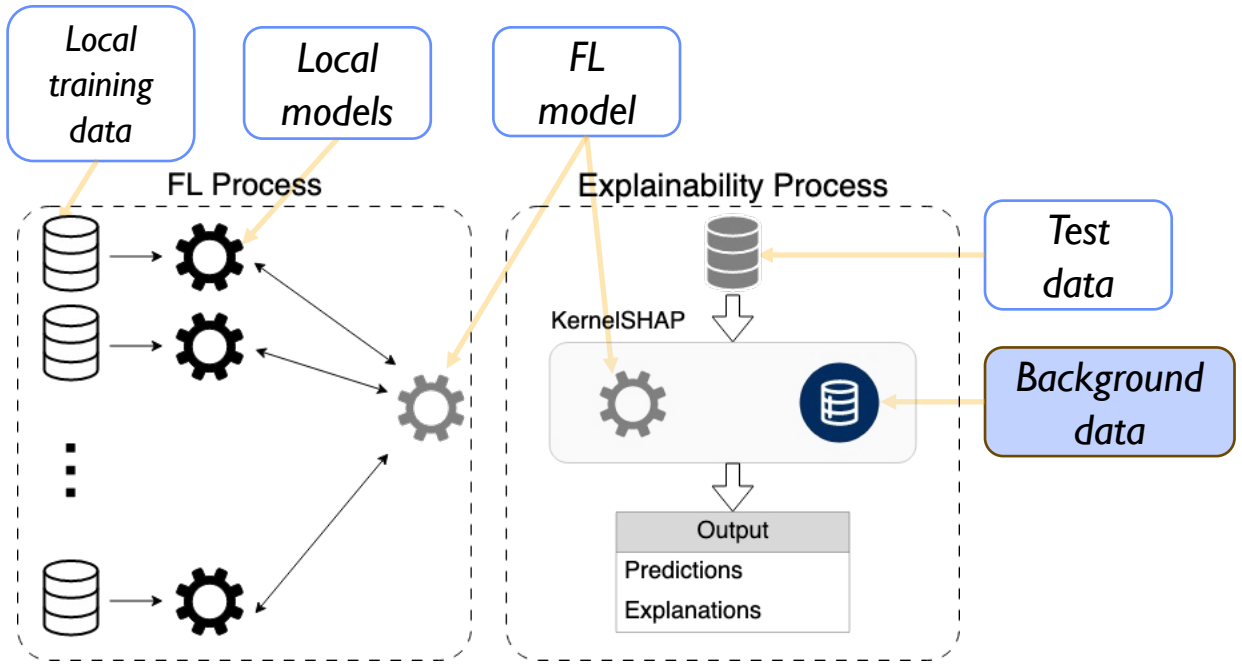
- The choice of the **background dataset impacts the resulting explanations**
- In the FL setting the **training set is not available in its entirety** to any party

Desiderata

- **Privacy preservation:** the explainability process should not violate privacy (as a constrain of the FL setting)
- **Consistency**: explanations of the same data instance for the FL model are identical for different participants
- **Accuracy:** explanations in FL match those that would be obtained in the traditional *centralized* setting

Federated SHAP – how to design a *proper* and *common* background dataset

- **Start** communication topology with **horizontally** partitioned data
- The model learned in a federation fashion is ***opaque*** (it requires post-hoc techniques)
- ***non-i.i.d.* setting**: local data follow distributions different from each other and from the overall distribution



Background dataset generation through Federated Fuzzy Clustering

- **Privacy preserving** summarization of scattered data
- **Cluster centers** are exploited as background
  - **common**, i.e., shared to all participants
  - **representative** of the entire data distribution
- Federated-FCM* is adopted but the choice of the clustering algorithm is not critical for our objective

*Corcuera Bárcena et al. ***A federated fuzzy c-means clustering algorithm.*** (2021)

Experimental setup – Baseline approaches

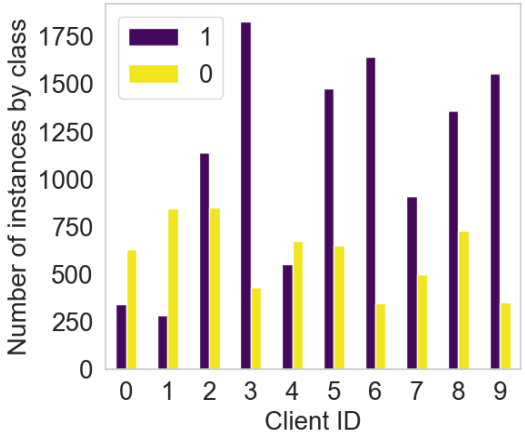
*BG = Background dataset	**Ensure consistency** (same background for all participants)	**Ensure accuracy** (represent the actual overall data distribution)	**Preserve privacy**
Federated SHAP **BG** ← K cluster centers obtained through Federated FCM	✅	✅	✅
Centralized **BG** ← union of the data locally stored in the clients	✅	✅	✖
Random **BG** ← randomly sampling K instances from a uniform distribution over the input space	✅	✖	✅
Local^m **BG^m** ← K cluster centers obtained through local FCM on the m-th participant	✖	✖	✅

Experimental setup

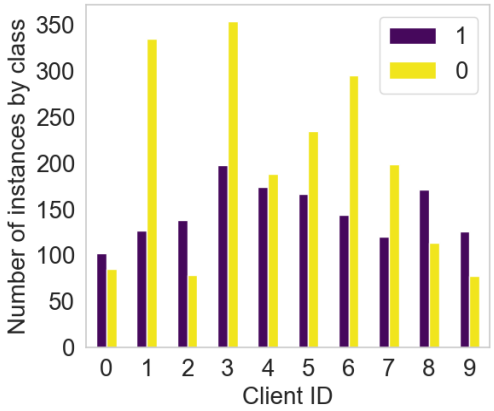
- ***Black box*** model: MLP-NN, two hidden layers each with 128 units
- **Ten clients:** *cross-silo* FL setup with horizontally partitioned data
- **Four datasets:** two for classification and two for regression
- ***non-iid*** scenario with both *quantity* skew and *label distribution* skew
- Unique, hold-out **test set** for each dataset (e.g., on the server)

TABLE II
DATASETS DESCRIPTION.

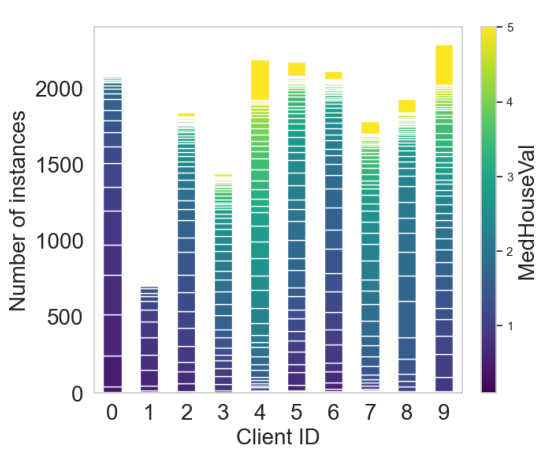
Dataset	Source	Task	N	N_{train}	N_{test}	F
Magic (MA)	[22]	C	19020	17118	1902	7
Rice (RI)	[22]	C	3810	3429	381	7
California (CA)	[23]	R	20640	18576	2064	8
Abalone (AB)	[22]	R	4177	3759	418	7



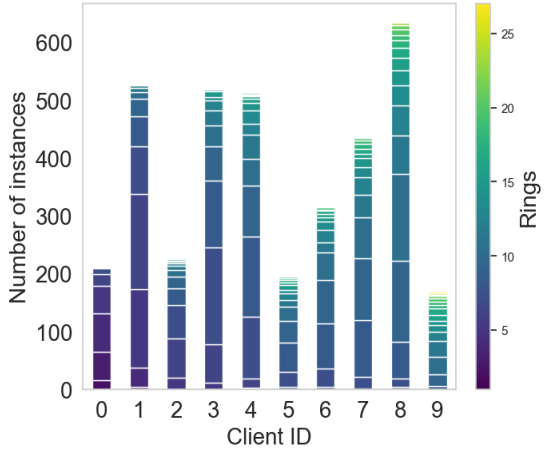
Magic



Rice



California

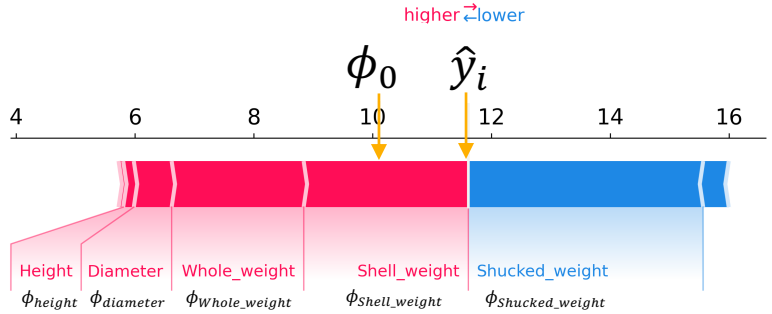


Abalone

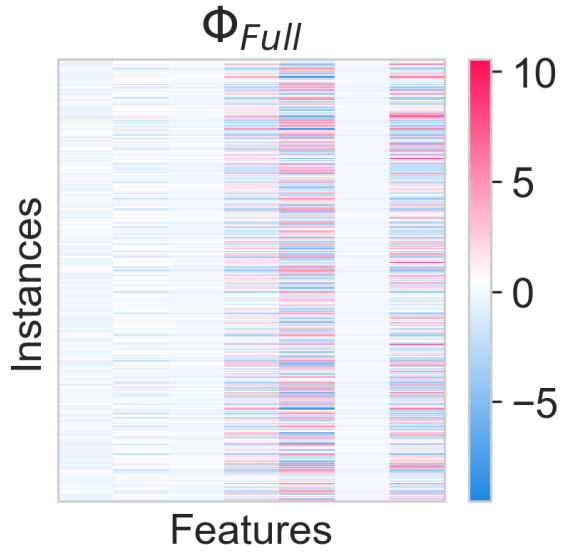
Experimental setup

From ***force plot:***
contribution of each feature j to the prediction for instance i

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{F} \phi_j$$



To ***heatmap of explanations:***
matrix $\boldsymbol{\phi}$ of Shapley Values for the test set



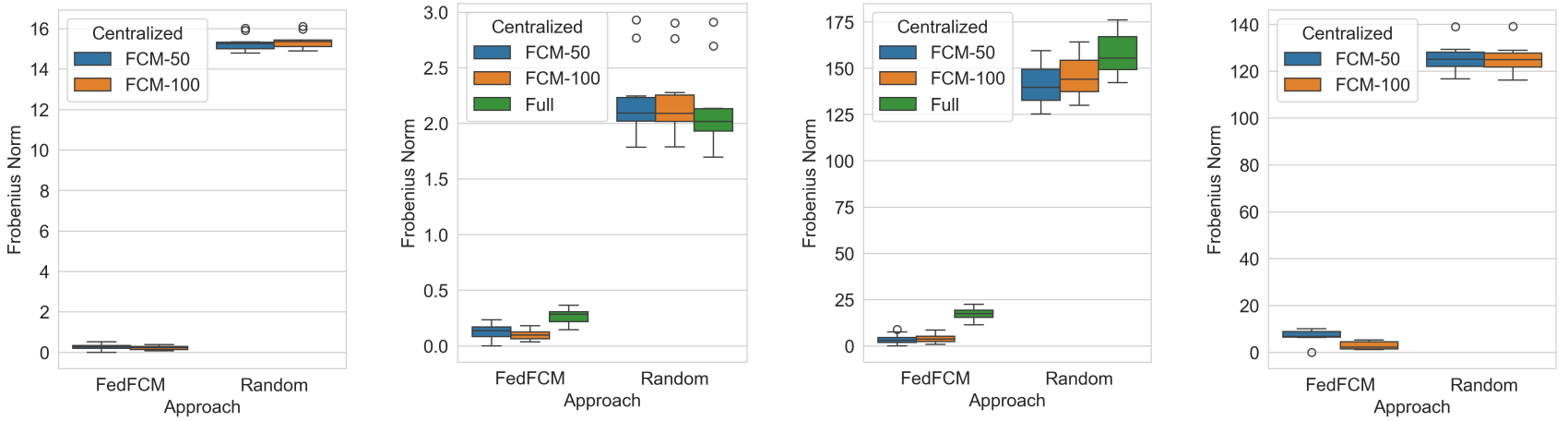
- **Comparison of approaches**

$$\|\boldsymbol{\phi}_A - \boldsymbol{\phi}_B\|_F = \sqrt{\sum_i \sum_j |\phi_A(i,j) - \phi_B(i,j)|^2}$$

- *e.g.:* A = *FederatedSHAP*, B = *Centralized*

Accuracy of explanations

- **Accuracy** $\stackrel{\text{def}}{=}$ explanations match those that would be obtained in the traditional *centralized* setting
- Three *centralized* versions
  - **BG** ← Full training
  - **BG** ← FCM, 50 centers
  - **BG** ← FCM,100 centers
- Ten values for each approach with different random seed

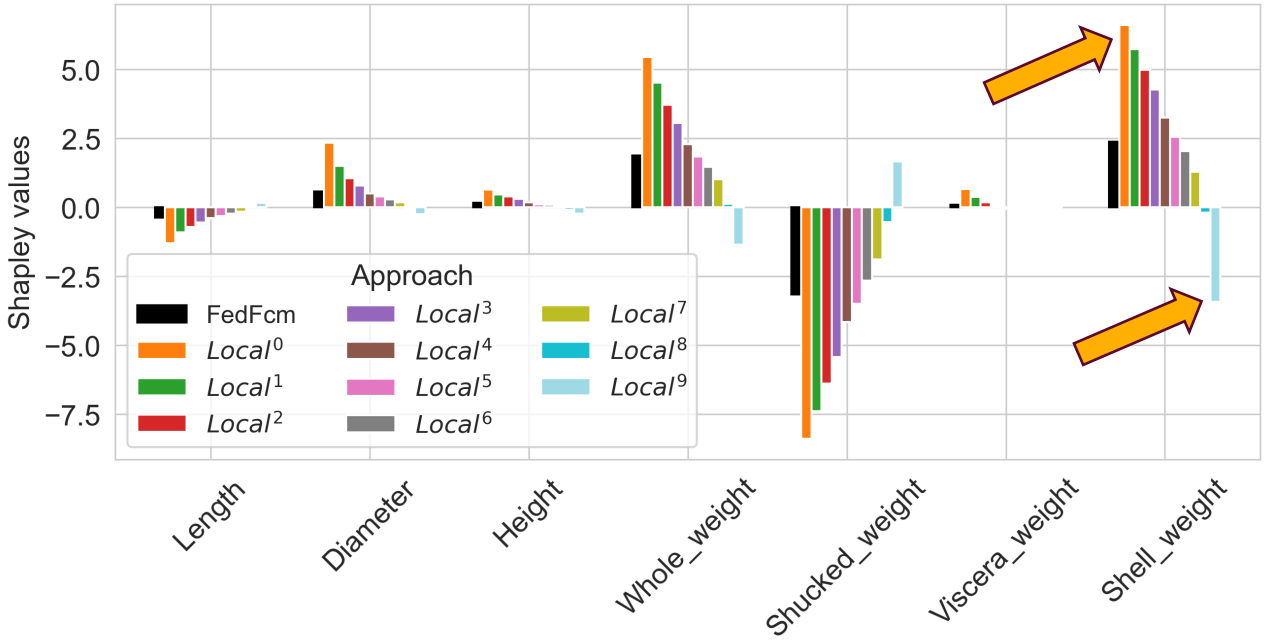


*Discrepancy of both the **Federated SHAP (FedFCM)** and the **Random** approach with the baseline centralized approaches in terms of Frobenius norm of the pairwise difference of ϕ matrices*

- **Federated SHAP:** low discrepancy with the *centralized* case, low variability
- **Random:** high discrepancy with the *centralized* case, high variability

Consistency of explanations

Consistency $\stackrel{\text{def}}{=}$ explanations of the same data instance for the FL model are identical for different participants



Shapley values for an instance of the Abalone dataset for each client

Local

- Background datasets (derived from local training sets) vary from client to client
- Particularly evident in non-i.i.d. settings
- **Consistency** not achieved: misalignment (i.e., variability) of client-side explanations

Federated SHAP, Centralized, Random

- Background datasets is unique
- Consistency achieved
- Black bars: Shapley values based on **Federated SHAP**

Conclusion

- **Federated SHAP**: approach for simultaneously addressing two requirements towards **trustworthy AI**
  - Privacy preservation → addressed through Federated Learning
  - Explainability → addressed through SHAP as *post-hoc* technique
- **Main goal**: obtaining accurate and consistent explanations in the federated setting, still ensuring privacy
- **Main challenge**: design a *proper* and *common* background dataset for the execution of SHAP
- **Key idea:**
  - Federated clustering procedure over scattered participants local data as a data summarization technique
  - Resulting cluster centers constitute the common *background dataset*
- **What's next:**
  - FL and other post-hoc explainability methods, possibly involving different data types (e.g., images and texts)

Thanks for your attention