GLOR-FLEX: Local to Global Rule-based EXplanations for Federated Learning

Rami Haffar
Department of Computer Engineering and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia,
Universitat Rovira i Virgili
Tarragona, Catalonia
rami.haffar@urv.cat

Francesca Naretto
KDDLab
University of Pisa
Pisa, Italy
francesca.naretto@unipi.it

David Sánchez
Department of Computer Engineering and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia,
Universitat Rovira i Virgili
Tarragona, Catalonia
david.sanchez@urv.cat

Anna Monreale
KDDLab
University of Pisa
Pisa, Italy
anna.monreale@unipi.it

Josep Domingo-Ferrer
Department of Computer Engineering and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia,
Universitat Rovira i Virgili
Tarragona, Catalonia
josep.domingo@urv.cat

2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) | 979-8-3503-1954-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/FUZZ-IEEE60900.2024.10611878

***Abstract*—The increasing spread of artificial intelligence applications has led to decentralized frameworks that foster collaborative model training among multiple entities. One of such frameworks is federated learning, which ensures data availability in client nodes without requiring the central server to retain any data. Nevertheless, similar to centralized neural networks, interpretability remains a challenge in understanding the predictions of these decentralized frameworks. The limited access to data on the server side further complicates the applicability of explainers in such frameworks. To address this challenge, we propose GLOR-FLEX, a framework designed to generate rule-based global explanations from local explainers. GLOR-FLEX ensures client privacy by preventing the sharing of actual data between the clients and the server. The proposed framework initiates the process by constructing local decision trees on each client's side to produce local explanations. Subsequently, by using rule extraction from these trees and strategically sorting and merging those rules, the server obtains a merged set of rules suitable to be used as a global explainer. We empirically evaluate the performance of GLOR-FLEX on three distinct tabular data sets, showing high fidelity scores between the explainers and both the local and global models. Our results support the effectiveness of GLOR-FLEX in generating accurate explanations that efficiently detect and explain the behavior of both local and global models.**

***Index Terms*—Explainable AI, TREPAN trees, federated learning, HOLDA, GLOCALX**

This research was partially supported by MCIN/AEI/10.13039/501100011033 (PhD grant PRE2019-089210 and project PID2021-123637NB-I00 "CURLING"), INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair), the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez, and grant 2021 SGR 00115), PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme, PNRR -"SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO, ERC-2018-ADG G.A. 834756 XAI: Science and technology for the eXplanation of AI decision making (https://xai-project.eu/index.html).

I. INTRODUCTION

Artificial intelligence (AI) has become a vital part of everyday life [1], especially since the blooming of deep learning (DL) [2], with applications ranging from personalized recommendations and voice assistants to image recognition, self-driving vehicles, and aid diagnostic systems. However, training robust AI models requires large amounts of data [3]. Unfortunately, gathering enough training data is not always feasible due to privacy and copyright regulations [4], which limit the ability of most individuals or organizations to train their specialized models independently.

Federated learning (FL) [5] has been proposed as a promising solution to the problem of training in case of data limitation. The idea is that multiple entities collaborate to train a global model without sharing their private data. Despite its success at training robust models, a main drawback of FL is that it only produces a static global model to be used and shared by all the participants, but it lacks the ability to deliver personalized local models adapted to the requirements of each specific participant [6].

To address this limitation, the Hierarchical crOss-siLo feDerated Averaging (HOLDA) method has been proposed [7].

HOLDA introduces a hierarchical approach that allows each participant to specialize their local models according to their local data and needs, thereby ensuring that the training will not only yield a robust global model, but also local models tailored to the unique requirements of each participant.

However, the inherent opacity of AI models, especially DL models, is a serious concern [8]. The vast majority of models, including those trained through federated learning (and HOLDA, in particular), generate predictions that are challenging for humans to comprehend. This lack of interpretability raises significant ethical and legal issues, since users cannot trust the predictions made by AI models without explanations [9]. Additionally, the enormous amount of data used for model training might contain some human biases in their annotations, which the model will inherit [10]. The need for explainability is beginning to show up in legal regulations and ethics guidelines, such as the General Data Protection Regulation (GDPR) [11], which states the right of citizens to an explanation of automated decisions affecting them, and the European Commission's Ethics Guidelines for Trustworthy AI [12] and the EU Artificial Intelligence Act [13], which insist on the organizations that make automated decisions to be prepared to explain them at the request of the affected citizens. Previous attempts to explain the predictions of DL models have been based on the assumption that the model owner possesses all the necessary elements to generate explanations. This assumption holds true primarily in centralized settings. However, in FL (and HOLDA), participants in the training process own the data, and the centralized server does not. As a result, the server is unable to build an explainer for the global model due to the absence of data on its side.

Contributions and Plan

Our work introduces a novel approach called Local to Global Rule-based eXplanations for Federated Learning (GLOR-FLEX) that explains the prediction of decentralized DL systems, with a particular focus on the HOLDA hierarchy. The proposed method aims to generate explanations of the global model predictions by leveraging insights extracted from individual participants in the training process while preserving the privacy of participants.

In our approach, TREPAN decision trees [14] are employed to create individualized local explainers for each training participant. These local explainers capture the behavior of each local model prediction. By using a Gaussian mixture model, each participant then creates synthetic data with a distribution similar to that of their real training data. These synthetic data instances are subsequently transmitted through the hierarchy to facilitate the construction of global explanations. The rules extracted from each TREPAN decision tree are also transmitted alongside the synthetic data to facilitate the generation of the global explanation. At the upper level of the HOLDA hierarchy, these rules are merged and generalized using the GLOCALX technique [15]. This innovative approach allows obtaining accurate global explanations originating from the local explainers without compromising the privacy of the training data owned by individual participants.

We conducted an empirical validation to demonstrate that the local explainers accurately mimic the behavior of local models on three distinct tabular data sets. Reported results show high fidelity scores between the explainers and both the local and global models. On the one hand, the rules extracted from the explainers reliably captured their behavior. On the other hand, on the server side, the merged rules appropriately reflected the behavior of the global model. Finally, the predictions made by the global model were accurately explained by using the merged rules.

The remainder of this paper is organized as follows. Section II summarizes previous research to explain the predictions of federated learning. Section III presents the background used by the proposed method. Section IV describes the GLOR-FLEX approach in detail. Section V reports and discusses experimental results. Finally, Section VI gathers the conclusions and depicts several lines of future research.

II. Related work

This paper addresses the problem of providing explanations in an FL setting. For this reason, in this section we briefly describe the state of the art in FL, followed by explainable AI (XAI) techniques and finally, we tackle the problem of XAI techniques for FL approaches.

a) ***FL****:* Neural networks (NN) exhibit good prediction performance with large datasets. When data reside across diverse parties, like mobile devices, centralized training is impractical due to privacy regulations (e.g., GDPR). In response, McMahan et al. proposed federated learning (FL) [5], enabling distributed parties to train a global model while safeguarding their data privacy. Since then, FL has been applied to diverse tasks, predominantly focusing on image and text data. FL has two approaches: *(i) cross-device*, involving stateless mobile devices (e.g., smartphones), and *(ii) cross-silo*, dealing with stateful organizations (e.g., hospitals, data centers) capable of saving and reusing intermediate training states. Further details on these scenarios are explored in [16]. The first algorithm in this context is federated averaging (FedAvg) [5]. In each iteration, the server receives the updated local models from the parties and subsequently aggregates these models to update the global model. In this work we exploit HOLDA [7], a cross-silo hierarchical FL approach in which the local clients share only the model that generalizes best on the local data and empirically showed best generalization capabilities. An in-depth description of HOLDA is given in Section III.

b) ***XAI****:* Explainability is now a crucial area of research in AI, especially for achieving trustworthy artificial intelligence. A comprehensive overview of interpretability techniques in machine learning (ML) is presented in [10], where two types of explanation models are identified: *global* and *local* explainers. The latter explain the prediction for individual instances [17]–[20], whereas the former explain the logic of the entire ML model [14], [21], [22]. We focus on global

explanations and in particular on TREPAN [14], presented in Section III.

A very important aspect of XAI techniques concerns the nature of the explanations they offer, such as rules, feature importance, or saliency maps. Given our interest in tabular data, we opt for rules as the preferred explanation. Rules mirror the logical reasoning of humans, involving premises and consequences, and they articulate the use of feature values within the dataset [10], [19], [23].

*c) **XAI in FL**:* Despite the longstanding popularity of FL and XAI, the task of explaining FL models has only recently gained attention. In [24] the authors proposed a survey of the XAI approaches tailored for FL models. Most of these methods offer post-hoc explanations through the assessment of feature importance [25], [26]. Given the potential privacy risks, Wang proposes a method to explain models based on Shapley values, that aims at striking a balance between interpretability and privacy [26]. The approach reveals detailed feature importances for owned features and provides a unified feature importance for features from other parties. Another work is [27], where Shapley values are employed in a horizontal FL architecture. In this scenario, local models generate explanations, and the global model merely aggregates client-side explanations, ensuring that no records of the training data are shared. In [25], Fiosina addresses interpretability challenges in horizontal FL, using FL to predict the taxi trip duration through the FedAvg algorithm. Integrated gradients are used to explain the model [28]. Lastly, an interesting work is also [29], in which the authors presented FED-XAI, a framework that employs inherently interpretable models in an FL setting. Notably, no other work has addressed the problem of providing explanations for neural-network based FL by exploiting rules as is our case.

III. Background

This section provides insights into the various techniques employed for generating explanations through the proposed approach.

A. The HOLDA Federated Learning Approach

We leverage HOLDA (Hierarchical crOss-siLo feDerated Averaging) [7] as our FL algorithm[1]. The main steps of the training process of HOLDA are depicted in Figure 1. This algorithm, which is tailored for tabular data, recursively trains a neural network NN in a hierarchical cross-silo FL domain, i.e., in a setting in which all the clients of the federation are stateful. The main objective of HOLDA is to endow NN with good generalization capabilities in all the nodes of the federation, both at the local and the global levels. In the remainder of this work, we consider the so-called *centralized* setting, in which we have a set $\mathcal{C}$ of clients directly connected to a server S, with no intermediate layers.

The process is started by the global server S, which randomly initializes the parameters w_g of its NN. At this point,

[1]The source code of HOLDA is available at https://github.com/michelefontana92/HOLDA

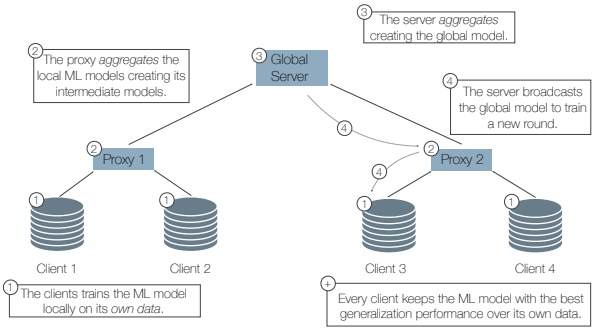

Fig. 1: Summary of the main training steps performed by HOLDA

the server S begins the global training phase, which consists of a loop of N global iterations. At each epoch $j = 1, \ldots, N$, S asks for model updating to its child nodes (that is, the clients) starting from the parameters w_g^j. Since in HOLDA the clients are stateful, we associate to each client $c \in \mathcal{C}$ an internal state $\sigma_c = \langle w_{best}, s_{best} \rangle$. Each client c trains the model with the received parameters on its own private data. If, during the local training, c finds a model w' having a validation score $s' > (\sigma_c \rightarrow s_{best})$, then it updates its internal state, by marking w' as the new best model. Formally, we have that $\sigma_c \leftarrow \langle w', s' \rangle$.

At the end of the local training, each client sends to the parent node the weights of the model stored in its state. After collecting the updated weights, S aggregates them into a new base model with parameters w_g^{j+1}. Finally, S *(i)* queries the child nodes to get the evaluation of their models on the validation set, *(ii)* computes the average evaluation, and *(iii)* selects the best model among the ones produced so far. We remark that if client c, after the evaluation of w_g^{j+1}, finds out that the model received from the server generalizes better than the model c has in its state, c saves w_g^{j+1} and its score into σ_c.

Once the N epochs are concluded, as a final step of personalization, HOLDA empowers the clients to fine-tune their final best model $\sigma_c = \langle w_{best}, s_{best} \rangle$ and obtain $\sigma_{fine-tuned} = \langle w_{fine-tuned}, s_{fine-tuned} \rangle$. If the performance on the validation set satisfies $s_{fine-tuned} > (\sigma_c \rightarrow s_{best})$, the client updates $\sigma_c := \sigma_{fine-tuned}$. Consequently, the internal state of each node contains the parameters of the model that *generalizes best* on its validation data. Thus, unlike other FL algorithms, where the final output is just the global model, HOLDA is able to train several models at the same time, i.e., it trains one model for each node of the federation.

B. GlocalX

GLOCALX (GLObal to loCAL eXplainer) is an explanation method that hierarchically merges local explanations into a global explanation by providing simple and faithful models. For the purposes of this work, we exploit GLOCALX to explain the behavior of the NN trained using HOLDA. Given a black-box b, an ML model whose internals are difficult to understand,

GLOCALX provides an explanation of the overall behavior of b. The GLOCALX process starts with a set of local post-hoc explanations $\mathcal{E} = \{e_1, \ldots, e_n\}$, where each e_i is a single explanation describing the reason behind the classification of a single record x. Each e_i is in the form of a *logical* rule, composed of a set of premises and a consequence, which is one of the target labels of b. Given the set E, GLOCALX iteratively merges the explanations and returns a single explanation.

More specifically, GLOCALX first sorts explanations by using a similarity function and a quality criterion. This sorting guides the subsequent merging of explanations. After the merging, the process checks if there is an improvement. If there is, the merging is added to the final set of rules. At the end, the rules are finally filtered by fidelity, i.e., GLOCALX selects the rules with the highest fidelity for each class.

For our purposes, we apply GLOCALX by shifting from merging local explanations to merging global explanations coming from local clients.

C. Trepan Decision Trees

In 1995, Craven and Shavlik introduced TREPAN [14], one of the pioneering efforts to define an explanation method for neural networks. Their seminal work addressed the inherent challenge of neural networks in terms of human comprehensibility. In particular, the authors tackled the task of creating a symbolic representation for trained neural networks. Despite the absence of a well-defined taxonomy for explainable artificial intelligence (XAI) at that time, they proposed TREPAN as a post-hoc, model-specific, and global explainer. The proposed explanation method takes the form of a surrogate model – specifically a decision tree–, designed to closely align with the underlying neural networks and hence yield surrogate models with high fidelity.

The core process of TREPAN involves an inductive learning paradigm where the target concept is the function embodied by the neural network. Given a neural network to be explained, denoted as b, the TREPAN model e is built by exploiting b as an oracle that can be queried at will. Moreover, the construction of the tree follows a best-first expansion strategy. The selection of the optimal node is determined by its potential to maximally enhance the fidelity with respect to the behavior of the neural network. The split selection process is also refined; unlike traditional decision tree learning where the number of training samples used for split selection decreases, TREPAN allows selecting the best split considering a user-specified minimum number of samples. In particular, during the split selection for a given node, the oracle possesses knowledge of all the previously chosen splits along the path from the root to that node. This information serves as constraints on the feature values. Additionally, TREPAN incorporates a stopping criterion consisting of two conditions. One condition is user-defined, representing a parameter that specifies the maximum allowable number of nodes. This parameter is crucial for enhancing comprehensibility. In the absence of a specified value, the second stopping condition is that a leaf node exclusively covers instances of a single class with high probability.

IV. Local to Global Rule-based eXplanations

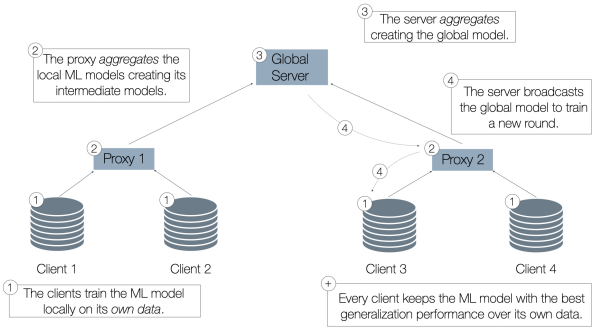

Fig. 2: Description of the proposed GLOR-FLEX approach. Given the FL structure of HOLDA, the explanations are extracted first from the local nodes, and then merged in the intermediate and global nodes.

In FL structures, and particularly in the HOLDA framework, the assumption is that central servers do not possess training data, which makes them unable to construct global explainers. In order to address this limitation and empower every participant in the NN training to elucidate predictions made by the NN models, we introduce a novel approach named GLOR-FLEX. Given the HOLDA structure, this innovative method employs TREPAN trees and GLOCALX to generate global explanations both at the client nodes and the global server. This enables each participant to gain insights into the behavior of their unique NN models, which makes predictions explainable at the client level while equipping the server to offer explanations about the behavior of the global model. These explanations are tailored for expert users with extensive experience in ML, who seek insights into the behavior of neural network models. The process by which GLOR-FLEX generates explanations is shown in Figure 2.

In line with the emphasis of the HOLDA hierarchy on securing the best model at each client, at each iteration clients selectively choose the model that performs best on their local data set, which is either the model trained in the current iteration or the previous best model. Moreover, clients retain the flexibility to fine-tune the final model, thus enhancing local model performance. As a result, each client possesses an individualized NN model, which in general diverges from the global model. To accommodate this divergence, GLOR-FLEX constructs a separate explainer at each client. This explainer comprises a TREPAN tree (see Section III) trained using the same data set that was used to train the NN model.

While the structure of the TREPAN, coupled with its extractable logical rules, provides insightful and interpretable explanations to each client regarding the behavior of their local models, the complete tree cannot be shared with the proxies or the server due to privacy considerations [30]. These concerns arise from the sensitivity of the local training data: sharing the TREPAN tree alongside the local model could potentially reveal additional information about the local training data.

Such a disclosure could render the system more vulnerable to privacy attacks, including membership inference attacks [31] and reconstruction attacks [32]. To ensure the privacy of the clients' data and mitigate potential risks, GLOR-FLEX has been designed to avoid the need for clients to share the TREPAN tree. To maximize client privacy, the proposed method requires the creation of a generalized tree achieved through regularization techniques [33], specifically by controlling the depth of the trees. The controlled depth not only enhances the model's performance on unseen data and prevents overfitting, but it also facilitates the generalization of rules extracted from the trees. This strategic approach aims to mitigate the risk of reconstructing the training data, thereby affording a robust privacy protection.

Upon completion of the training of its local generalized TREPAN tree explainer, each client proceeds to extract logical rules from it [34]. Extracting rules from a decision tree involves creating one rule for each path from the root to a leaf node. Along a given path, each splitting criterion is joined with a logical *and* to formulate the rule's antecedent ("IF" part). The class prediction held by the leaf node becomes the rule's consequent ("THEN" part). Subsequently, these rules are shared with upper-level entities in the HOLDA structure, in our setting the server.

For the server to sort and merge the received rules from the various clients, it is essential to evaluate the performance of the rules on test data. As previously highlighted, the server lacks direct access to any data. To overcome this challenge, we have opted for a solution wherein synthetic data sets accompany the rules. These data sets are generated by each client to aid the transition from local to global explanations. The creation of the synthetic data involves fitting a Gaussian mixture model (GMM) [35] to capture the distribution of the original local training data of each client. The GMM is a statistical model defined as

$$p(x) = \sum_{i=1}^{k} \pi_i \cdot N(x|\mu_i, \Sigma_i),$$

where k is the number of components of the mixture, π_i is the weight of the i-th component, and $N(x|\mu_i, \Sigma_i)$ is the multivariate Gaussian distribution with mean μ_i and covariance matrix Σ_i. By sampling the fitted GMM [36], the client produces entirely synthetic data while preserving the distribution characteristics of its original local data set.

After receiving the logical rules and synthetic data sets from the clients, the server, in turn, utilizes the GLOCALX framework (see Section III) to assess the importance of these rules and merges them. This process yields a representative set of rules capable of providing comprehensive explanations for the predictions made by the global NN model.

V. EXPERIMENTS

In this section we report the experiments conducted to validate GLOR-FLEX[2].

[2]All the codes for the experiments are available at https://github.com/anonymous16534/GLOR-FLEX

All experiments were conducted on a machine equipped with an AMD Ryzen 5 3600 CPU (base speed 3.6 GHz), and 32 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM).

A. Data Sets

As mentioned above, our focus is on tabular data. We considered three data sets with different statistical distributions:

- ADULT data set [37]. This is a *de facto* standard data set[3]. ADULT contains $48,842$ records of census income information, with 14 numerical and categorical attributes. For each categorical attribute, we re-coded categories as numbers to obtain a numerical version of the attribute.
- "PAMAP2 Physical Activity Monitoring" (a.k.a. ACTIVITY) data set [38][4]. This data set contains continuous measurements of 3 inertial body sensors and a heart-rate monitor worn by 9 subjects who performed 18 different activities such as walking, watching TV, etc. First, as recommended by the releasers of the data set, we discarded the transient activity (e.g., going to the next activity location). Second, we mapped the various types of activity into two categories indicating whether the activity involved displacement or not (e.g., walking was mapped to "displacement" and watching TV to "not displacement"). We obtained a data set of $1,942,872$ records, of which $1,136,540$ records were labeled as "displacement" and $806,332$ as "not displacement". Each record contained 54 numerical attributes corresponding to timestamp, heart rate, and 17 sensor data feeds. The classification task consisted in detecting whether the subject was performing an activity involving physical displacement.
- GAUSSIAN data set. This is a synthetic binary classification data set consisting of $40,000$ records, featuring 30 numerical attributes. It was generated by using a GMM statistical model. Given the controlled generation of these data, no pre-processing was needed. The decision to employ this synthetic data set stems from our intention to execute an experiment within a controlled environment. This choice allows us to manipulate attributes precisely and ensures a standardized setting for rigorous testing.

B. HOLDA Training

After preparing the data sets, we focused on the training of the HOLDA models. In order to avoid having too small local data sets, we considered 4 client nodes and 1 server. Given the inherent decentralized nature of HOLDA, we considered two settings: (i) *iid* setting, in which we distributed the data randomly and equally among all clients, ensuring that no records were repeated between any two different clients; (ii) *non-iid* setting, in which we deliberately introduced an imbalance by allocating 80% of records from one class and 20% from the remaining class to individual clients. We proceeded

[3]https://archive.ics.uci.edu/dataset/2/adult

[4]https://archive.ics.uci.edu/dataset/231/pamap2+physical+activity+monitoring

to train all the different NN models for 10 global epochs. Each global epoch comprised 3 local training epochs, with each client completing the training by fine-tuning the local model for an additional 10 local epochs. We also compared the performance of the HOLDA models with that of centralized networks with the same structure trained for 30 epochs. This approach resulted in exceptionally good performance across all data sets, as reported in Table I[5]. In this experiment, each client employed 70% of its local data as training data, and the remaining 30% as testing data to evaluate the local model's performance.

The best results in both settings were obtained with the GAUSSIAN data set, in line with the results in the centralized settings, due to the controlled creation of the synthetic data. Very high performance was reached in all nodes for ACTIVITY in both settings, which is consistent with the centralized settings. However, although the converged global model exhibited similar high performance as the local models in the *iid* settings, in the *non-iid* settings it stayed below the local models, with a global model accuracy of 68%. This can be primarily attributed to the substantial divergence among the local models, which makes it difficult to reach similar performance. In both settings, slightly lower prediction performance was obtained for ADULT. However, the results of ADULT are aligned with those in centralized settings and the state of the art on this data set [42]. Hence, we can claim that we achieved top prediction performance for all data sets in both *iid* and *non-iid* settings.

C. Explanation Generation

Once the NN models were trained within the HOLDA framework, we proceeded with the generation of the explanations. We remark that our objective was to craft global explanations that delineate the overall behavior of the model.

The explanation generation process started at the client nodes. Each client trained a TREPAN decision tree on its local data. We validated the TREPAN models by means of fidelity and accuracy, as presented in Table II. Fidelity is the number of the matching predictions between the NN model and the explainer divided by the total number of predictions. Ensuring high fidelity is of paramount importance: it serves as a key metric in the XAI field, to validate the consistency between the predictions of the explanation model and those of the NN model under analysis. In Table II we find consistently high fidelity values for all models in both settings. Thus, TREPAN explanation models faithfully replicate the behavior of the original local NN, thereby giving confidence in the provided explanations.

A comparative analysis of the results in Table I on the performance of HOLDA and those in Table II on the performance of TREPAN reveals a direct correlation between the goodness of the explanation model and the prediction performance of the NN models. In particular, for GAUSSIAN and ACTIVITY, where the local NN achieves very high performance, we observe both high fidelity and accuracy in the explanation models. We can also notice that the depth of the trees is not large (at most 15, reached in the case of the ACTIVITY data set, the one with the highest number of attributes), which ensures effective generalization. Note also that rules involving a reduced set of attributes have been shown to be more *comprehensible* [43].

Based on the trained TREPAN models, the actual final global explanation model is generated by the server by merging the results of the clients. To that end, rules are extracted from the local TREPAN trees. In this way, we can exploit a state-of-the-art methodology, GLOCALX, to merge the explanations and hence obtain a global explanation composed of the most important rules. To sort and merge the rules, GLOCALX requires data for evaluating the importance of each rule. To cater for this, each client sampled 1000 synthetic records using the GMM model and sent them to the server.

The outcomes obtained from the application of GLOCALX are reported in Table III. This table incorporates a parameter, α, which has significant importance within the GLOCALX framework, as it enables the selection of rules based on their fidelity. A higher value of α means higher fidelity and corresponds to a reduced number of rules in the final set. In other words, as the α parameter increases, fidelity grows, whereas the number of rules decreases. For the sake of comprehensiveness, we present results for various values of α. However, to establish a definitive global set of rules, we recommend a minimum value of $\alpha = 80$. From our experiments, this choice ensures that all data sets achieve a fidelity of at least 77%.

Table III also incorporates additional metrics, such as coverage, which represents the fraction of records satisfying the antecedent of at least one rule. Notably, all experiments demonstrate high coverage. Furthermore, the table provides information on the number of rules, categorized by class (0 or 1), and their average length. It is noteworthy that the average rule length is relatively short, with the average maximum length reaching around 10 for the ACTIVITY data set in the *iid* settings, which is the data set with the largest number of variables. Whereas the number of rules may seem very large, it is essential to realize that we aim to explain the overall behavior of a complex NN model. Excessively restricting the number of rules would compromise the effectiveness of the explanation model.

An alternative solution would involve avoiding the entire GLOR-FLEX process and allowing the server to train the TREPAN tree using synthetic datasets generated by various clients. This approach lacks personalized explanations for individual clients, which may be undesirable in cross-silo FL where client nodes are designed to function autonomously. However, it has the advantage of speeding up the process by training a single TREPAN tree, thereby reducing the information-sharing overhead. We conducted an experiment in the *iid* setting to explore this strategy. The results are summarized in Table IV. To ensure a fair comparison, we maintained the same depth for TREPAN indicated in Table III. The outcomes reveal

[5]The architecture of the NN is defined as $(\text{input features}, 256, 256, 2)$, following the configuration used in [39]. The model employs the cross-entropy loss function, the Adam optimizer [40], the ReLU activation function [41], a batch size of 512, and a learning rate of 0.005.

TABLE I: Performance of the NN of HOLDA. C_i refers to the i^{th} client's local model and *Global* refers to the global NN model in both the *iid* and *non-iid* settings. For comparisons, the C setting presents the results of the **centralized** setting, while the NA stands for Non Applicable.

		ADULT					ACTIVITY					GAUSSIAN				
		C_0	C_1	C_2	C_3	*Global*	C_0	C_1	C_2	C_3	*Global*	C_0	C_1	C_2	C_3	*Global*
C	**Loss**	NA	NA	NA	NA	0.29	NA	NA	NA	NA	0.00	NA	NA	NA	NA	0.00
	Accuracy	NA	NA	NA	NA	0.83	NA	NA	NA	NA	0.99	NA	NA	NA	NA	0.99
	Precision	NA	NA	NA	NA	0.86	NA	NA	NA	NA	0.99	NA	NA	NA	NA	0.99
	Recall	NA	NA	NA	NA	0.81	NA	NA	NA	NA	0.99	NA	NA	NA	NA	0.99
	F_1	NA	NA	NA	NA	0.82	NA	NA	NA	NA	0.99	NA	NA	NA	NA	0.99
IID	**Loss**	0.35	0.36	0.36	0.37	0.38	0.01	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	Accuracy	0.84	0.82	0.83	0.81	0.80	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1
	Precision	0.85	0.84	0.86	0.83	0.85	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1
	Recall	0.84	0.82	0.83	0.81	0.80	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1
	F_1	0.85	0.83	0.84	0.82	0.81	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1
NON-IID	**Loss**	0.33	0.37	0.33	0.38	0.36	0.06	0.08	0.04	0.05	1.02	0.00	0.01	0.02	0.02	0.00
	Accuracy	0.83	0.82	0.83	0.81	0.81	0.99	0.99	0.99	0.99	0.68	0.99	0.99	0.99	0.99	0.99
	Precision	0.84	0.84	0.86	0.82	0.84	0.99	0.99	0.99	0.99	0.88	0.99	0.99	0.99	0.99	0.99
	Recall	0.83	0.82	0.83	0.81	0.81	0.99	0.99	0.99	0.99	0.68	0.99	0.99	0.99	0.99	0.99
	F_1	0.84	0.82	0.84	0.81	0.82	0.99	0.99	0.99	0.99	0.69	0.99	0.99	0.99	0.99	0.99

TABLE II: Depth, fidelity, and accuracy of TREPAN trees trained as local explanations for each client node in the *iid* and *non-iid* settings

		ADULT				ACTIVITY				GAUSSIAN			
		C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3
IID	**Max_Depth**	5	5	5	5	14	14	12	15	5	5	5	5
	Fidelity	0.89	0.89	0.91	0.90	0.88	0.89	0.88	0.90	0.96	0.96	0.97	0.96
	Accuracy	0.83	0.81	0.78	0.76	0.93	0.95	0.94	0.95	0.97	0.97	0.98	0.97
NON-IID	**Max_Depth**	5	5	5	5	13	13	13	13	5	5	5	5
	Fidelity	0.88	0.90	0.99	0.89	0.99	0.98	0.99	0.99	0.96	0.97	0.97	0.97
	Accuracy	0.80	0.82	0.81	0.83	0.99	0.99	1	1	0.97	0.97	0.98	0.98

lower performance in terms of fidelity and coverage compared to GLOR-FLEX, with the exception of the ACTIVITY dataset, where both methods perform similarly. In this case, it is worth noting that the rules extracted are longer on average (13 vs 10 rule length). This empirical analysis demonstrates the superiority of our approach at generating high-performance explanation rules. Moreover, it provides the added benefit of producing personalized explanations for each client.

Finally, in Table V, we present the execution times of the various steps of our procedure. The results are consistent with state-of-the-art approaches in XAI. Furthermore, it is worth noting that, to generate global explanations, the process only needs to be carried out once, as an added step to the training time.

VI. Conclusion

We have presented GLOR-FLEX, a framework that generates local to global rule-based explanations for FL. Our approach leverages TREPAN decision trees to generate explanations at each participating client. By merging rules derived from these local trees under GLOCALX, that strategically sorts and merges these rules, we establish a comprehensive set of rules suitable for a global explainer on the server side. GLOR-FLEX addresses the inherent data limitations on the server side, which are common in decentralized learning frameworks such as FL. This enhances model interpretability both at the local and the global levels.

Empirical results show high fidelity for the TREPAN trees across three different data sets, both for *iid* and *non-iid* settings. This attests to their ability to provide accurate and comprehensive explanations for the local NN models. Moreover, our experiments demonstrate that the merged rules exhibit both high fidelity and coverage, thereby resulting in accurate explanations for the global model. In particular, GLOR-FLEX effectively addresses the lack of data on the server side while preserving privacy for the local data of clients. This is made possible by the transformation from local to global explanations being seamlessly executed by GLOR-FLEX without requiring any sharing of real data between the clients and server.

As future work, we plan to explore rule selection mechanisms based on user preferences, which should allow delivering customized explanations. Furthermore, a study on the privacy risks associated with sharing the local explainer alongside updates for the NN model will be conducted. Finally, we aim to conduct a user study to assess the quality of the explanations provided.

References

[1] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation*, vol. 2, no. 4, 2021.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] N. M. Jebreel, J. Domingo-Ferrer, and Y. Li, "Defending against backdoor attacks by layer-wise feature analysis," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2023, pp. 428–440.

[4] B. Otto, "Quality and value of the data resource in large enterprises," *Information Systems Management*, vol. 32, no. 3, pp. 234–251, 2015.

TABLE III: Experimental results on the application of TREPAN and GLOCALX to explain HOLDA: fidelity, coverage, total number of rules, number of rules per class, average length of rules per class.

Dataset		α	0	10	20	30	40	50	60	70	80	90	100
ADULT	IID	**Fidelity**	0.76	0.76	0.76	0.76	0.76	0.77	0.80	0.83	0.85	0.88	0.74
		Coverage	1	1	1	1	1	1	1	1	1	0.99	0.99
		# Rules	1761	1710	1657	1603	1534	1397	1304	1235	1174	1117	1067
		# Rules C_0	1066	1066	1066	1066	1066	1066	1066	1066	1066	1066	1066
		# Rules C_1	695	644	591	537	468	331	238	169	108	51	1
		Avg. Rules C_0	6.36	6.36	6.36	6.36	6.36	6.37	6.36	6.36	6.36	6.36	6.36
		Avg. Rules C_1	6.87	6.86	6.85	6.84	6.81	8.86	6.88	6.82	6.82	6.69	6
	NON-IID	**Fidelity**	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.80	0.78
		Coverage	1	1	1	1	1	0.99	0.99	0.99	0.99	0.98	0.98
		# Rules	118	113	108	103	98	85	80	75	70	65	61
		# Rules C_0	60	60	60	60	60	60	60	60	60	60	60
		# Rules C_1	58	53	48	43	38	25	20	15	10	5	1
		Avg. Rules C_0	4.29	4.28	4.28	4.28	4.28	4.28	4.28	4.28	4.28	4.28	4.28
		Avg. Rules C_1	4.26	4.28	4.30	4.27	4.23	4.36	4.30	4.20	4.10	4	5
ACTIVITY	IID	**Fidelity**	0.67	0.67	0.67	0.67	0.67	0.70	0.70	0.73	0.77	0.79	0.69
		Coverage	1	1	1	1	1	1	1	1	1	1	0.99
		# Rules	6194	6157	6114	6028	5336	4423	4328	4269	4223	4182	4146
		# Rules C_0	4145	4145	4145	4145	4145	4145	4145	4145	4145	4145	4145
		# Rules C_1	2049	2012	1969	1883	1191	278	183	124	78	37	1
		Avg. Rules C_0	9.75	9.75	9.75	9.75	9.75	9.75	9.75	9,750543	9.75	9.75	9.75
		Avg. Rules C_1	9.97	9.98	9.97	10.01	9.95	10.29	10.16	10.31	10.19	9.94	10
	NON-IID	**Fidelity**	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.78	0.76
		Coverage	1	1	1	1	1	1	1	1	1	1	1
		# Rules	2459	2427	2396	2360	2322	2267	2205	2053	1452	1387	1349
		# Rules C_0	1348	1348	1348	1348	1348	1348	1348	1348	1348	1348	1348
		# Rules C_1	1111	1079	1048	1012	974	919	857	705	104	39	1
		Avg. Rules C_0	8.25	8.24	8.25	8.25	8.25	8.25	8.25	8.25	8.25	8.25	8.25
		Avg. Rules C_1	8.24	8.28	8.33	8.35	8.39	8.44	8.51	8.57	8.07	8.15	8
GAUSSIAN	IID	**Fidelity**	0.93	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.96	0.97	0.90
		Coverage	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.96
		# Rules	108	103	97	92	86	81	76	71	66	61	56
		# Rules C_0	55	55	55	55	55	55	55	55	55	55	55
		# Rules C_1	53	48	42	37	31	26	21	16	11	6	1
		Avg. Rules C_0	4.45	4.45	4.45	4.45	4.45	4.45	4.45	4.45	4.45	4.45	4.45
		Avg. Rules C_1	4.58	4.62	4.60	4.65	4.70	4.73	4.66	4.47	4.63	4.66	5
	NON-IID	**Fidelity**	0.88	0.88	0.88	0.88	0.89	0.90	0.90	0.91	0.92	0.93	0.94
		Coverage	1	1	1	1	1	1	1	1	0.99	0.99	0.99
		# Rules	123	117	112	107	100	95	90	84	79	74	669
		# Rules C_0	68	68	68	68	68	68	68	68	68	68	68
		# Rules C_1	55	49	44	39	32	27	22	16	11	6	1
		Avg. Rules C_0	4.54	4.54	4.54	4.54	4.54	4.54	4.54	4.54	4.54	4.54	4.54
		Avg. Rules C_1	4.65	4.50	4.63	4.61	4.65	4.66	4.68	4.62	4.45	4.50	5

TABLE IV: Metrics for the TREPAN explainers trained on the server's side using synthetic data sets

	ADULT	ACTIVITY	GAUSSIAN
Max-depth	5	14	5
Fidelity	0.64	0.69	0.88
Coverage	0.94	0.87	0.89
# Rules	31	27	31
# Rules C0	18	20	14
# Rules C1	13	7	17
Avg. Rules C0	4.16	13.55	4.71
Avg. Rules C1	4.46	13.57	5.00

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[6] M. Fontana, F. Naretto, A. Monreale, and F. Giannotti, "Monitoring fairness in holda," in *HHAI2022: Augmenting Human Intellect*. IOS Press, 2022, pp. 246–248.

[7] M. Fontana, F. Naretto, and A. Monreale, "A new approach for cross-silo federated learning and its privacy risks," in *2021 18th International Conference on Privacy, Security and Trust (PST)*. IEEE, 2021, pp. 1–10.

[8] M. Khalili, "Against the opacity, and for a qualitative understanding, of artificially intelligent technologies," *AI and Ethics*, pp. 1–9, 2023.

[9] A. Brennen, "What do people really want when they say they want" explainable ai?" we asked 60 stakeholders." in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7.

[10] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, 2023.

[11] GDPR, "General Data Protection Regulation, Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016," *Official Journal of the European Union*, 2016.

[12] N. A. Smuha, "The eu approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.

[13] EU-AI-Act, "Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024," 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[14] M. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *NIPS*, 1995, pp. 24–30.

[15] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "Glocalx-from local to global explanations of black box ai models," *Artificial Intelligence*, vol. 294, p. 103457, 2021.

TABLE V: Execution time of GLOR-FLEX

		ADULT	ACTIVITY	GAUSSIAN
	Building TREPAN	680.06 s	2129.46 s	1465.05 s
Client side	**Training GMM models and generating the synthetic data**	0.21 s	1.93 s	0.32 s
	Generating the rules	13.41 s	14.59 s	0.32 s
Server side	**GLOCALX**	38.89 s	9869 s	42.28 s

[16] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, pp. 3347–3366, 2023.

[17] M. T. Ribeiro *et al.*, ""Why should I trust you?": Explaining the predictions of any classifier," in *ACM SIGKDD*, 2016, pp. 1135–1144.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017, pp. 4765–4774.

[19] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, and F. Giannotti, "Stable and actionable explanations of black-box models through factual and counterfactual rules," in *Data Mining and Knowledge Discovery*, Springer, Ed., 2022.

[20] R. Haffar, D. Sanchez, and J. Domingo-Ferrer, "Explaining predictions and attacks in federated learning via random forests," *Applied Intelligence*, vol. 53, no. 1, pp. 169–185, 2023.

[21] M. W. Craven and J. W. Shavlik, "Using sampling and queries to extract rules from trained neural networks," in *JMLR*. Elsevier, 1994, pp. 37–45.

[22] H. Deng, "Interpreting tree ensembles with intrees," *Int. Journal Data Science and Analytics*, vol. 7, no. 4, pp. 277–287, 2019.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations." in *AAAI*, vol. 18, 2018, pp. 1527–1535.

[24] J. L. Corcuera Bárcena, D. Mattia, P. Ducange, F. Marcelloni, A. Renda, F. Ruffini, and A. Schiavo, "Fed-xai: federated learning of explainable artificial intelligence models." *XAI.it@AI*IA, CEUR Workshop Proceedings*, 2022.

[25] J. Fiosina, "Explainable federated learning for taxi travel time prediction." *VEHITS. SCITEPRESS*, 2021.

[26] Z. Z. Guan Wang, Charlie Xiaoqian Dang, "Measure contribution of participants in federated learning." *2019 IEEE International Conference on Big Data (Big Data)*, 2019.

[27] A. M. Luca Corbucci, Riccardo Guidotti, "Explaining black-boxes in federated learning." *Longo, L. (eds) Explainable Artificial Intelligence. xAI 2023.*, 2023.

[28] L. M. Dominik Janzing and P. Blobaum, "Feature relevance quantification in explainable ai: a causal problem." *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

[29] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, and L. G. Baltar, "Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking." *Information 2022*, 2022.

[30] Z. Zhu and W. Du, "Understanding privacy risk of publishing decision trees," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2010, pp. 33–48.

[31] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[32] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1138–1156.

[33] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer, "Generalized and scalable optimal sparse decision trees," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6150–6160.

[34] J. de Oña, G. López, and J. Abellán, "Extracting decision rules from police accident reports through decision trees," *Accident Analysis & Prevention*, vol. 50, pp. 1151–1160, 2013.

[35] D. A. Reynolds *et al.*, "Gaussian mixture models." *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[36] C. Chokwitthaya, Y. Zhu, S. Mukhopadhyay, and A. Jafari, "Applying the gaussian mixture model to generate large synthetic data from a small data set," in *Construction Research Congress 2020*. American Society of Civil Engineers Reston, VA, 2020, pp. 1251–1260.

[37] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[38] A. Reiss, "PAMAP2 Physical Activity Monitoring," UCI Machine Learning Repository, 2012, DOI: https://doi.org/10.24432/C5NW2H.

[39] M. Fontana, F. Naretto, A. Monreale, and F. Giannotti, "Monitoring fairness in HOLDA," in *HHAI 2022: Augmenting Human Intellect - Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence*, vol. 354. IOS Press, 2022, pp. 246–248. [Online]. Available: https://doi.org/10.3233/FAIA220205

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[42] N. Chakrabarty and S. Biswas, "A statistical approach to adult census income level prediction," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2018, pp. 207–212.

[43] J. Stecher, F. Janssen, and J. Fürnkranz, "Shorter rules are better, aren't they?" in *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*. Springer, 2016, pp. 279–294.