Security Threats to Explainable Classifiers in Federated Learning

Mattia Daole* ‡, Pietro Ducange*, Francisco Herrera†, Francesco Marcelloni*,
Alessandro Renda*, Nuria Rodríguez-Barroso†

* Dept. of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
Email: mattia.daole@phd.unipi.it, {pietro.ducange, francesco.marcelloni, alessandro.renda}@unipi.it

‡ Dept. of Informatics, University of Pisa, Largo Bruno Pontecorvo, 3, 56127 Pisa, Italy

† Dept. of Computer Science and Artificial Intelligence, DaSCI, University of Granada, Spain
Email:herrera@decsai.ugr.es, rbnuria@ugr.es

Abstract—**The decentralized nature of federated learning (FL) poses critical challenges related to security: Clients participating in the process may not necessarily be trustworthy and could engage in adversarial attacks, potentially undermining the integrity and reliability of the global machine learning model. Security concerns have been extensively investigated in traditional FL, where collaboratively learned models are typically deep neural networks. However, this class of models does not meet the requirement of explainability, which is considered essential for the trustworthiness of AI systems. In this work, we present an analysis on security threats to FL of explainable models, namely fuzzy rule-based classifiers (FRBCs). We outline the types of attacks a malicious client may implement, and assess, through a preliminary experimental analysis, the impact they have on FL of FRBCs in terms of global model performance. We also compare these findings with the effects of the same or similar well-established attacks in traditional FL of neural network models. Finally, we provide insights to improve the security of FRBCs learned in a federated fashion.**

Index Terms—**Federated Learning, Explainable Artificial Intelligence, Fuzzy Rule-based Classifiers, Security**

I. Introduction

Users, companies, institutions, and government entities are posing increasing attention towards trustworthiness in Artificial Intelligence (AI), given its evident pervasiveness in our daily lives. The European Commission (EC), for example, enacted the “AI Act” in June 2024 (online artificialintelligenceact.eu), thus introducing the first common regulatory and legal framework for AI in Europe. The document partly builds upon a previous initiative documented in the “Ethics Guidelines for Trustworthy AI” [1]: a high-level expert group on AI established by the EC identified the requirements an AI system must meet to achieve trustworthiness, which are lawful, ethical, and robust. Among technical requirements, *transparency* and *privacy* play a crucial role.

The ability to understand *how* a model works and *why* decisions are made are fundamental aspects of transparency in AI and at the core of a branch named “explainable AI” (XAI) [2], [3]. Models like decision trees and rule-based systems are typically referred to as interpretable by-design, as they can be traced back to collections of rules in the form “IF *antecedent* THEN *consequent*”. Hence, the inference process turns out to be highly understandable to human observers. Other classes of models, such as Neural Networks (NN) and Deep Learning (DL), entail an inherent complexity that prevents an immediate understanding of their behavior: for this reason, they are typically referred to as black-box models and require adoption of post hoc techniques to explain their predictions.

The growing need for explainable AI intersects with the challenges posed by the privacy requirement. With the increasing number of widespread smart and connected devices, the traditional paradigm of centralized machine learning (ML) suffers from a fundamental problem: data owners are reluctant to share their data assets for privacy concerns, and this hampers the training of models that are typically data hungry. Similar considerations apply to institutions and companies that handle sensitive information, such as hospitals or banks, as their data is organized in isolated data silos and cannot be shared with other parties. To overcome this challenge, Federated Learning (FL) [4], [5] has recently been proposed as a paradigm to enable collaborative learning of ML models. In FL, the learning objective is achieved by aggregating model updates or statistics computed locally by participating data owners: this approach eliminates the need to gather data at a central location for training an ML model. The most popular implementations of FL approaches involve NN models and stem from a protocol known as Federated Averaging (FedAvg) which allows the standard optimization algorithm (stochastic gradient descent, SGD) to be performed in a decentralized setting. As a consequence, efforts to design AI systems that comply with the privacy requirement have mainly addressed

This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” and the PNRR “Tuscany Health Ecosystem” (THE) (Ecosistemi dell’Innovazione) - Spoke 6 - Precision Medicine & Personalized Healthcare (CUP I53C22000780001) under the NextGeneration EU programme, and by the Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence). This research results from the Strategic Project IAFER-Cib (C074/23), as a result of the collaboration agreement signed between the National Institute of Cybersecurity (INCIBE) and the University of Granada. This initiative is carried out within the framework of the Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation).

the FL of black-box models, thus overlooking the explainability requirement.

Several works recently attempted to simultaneously meet the requirements of privacy and explainability, and they fall within a research topic named Fed-XAI [6], [7], acronym for FEDerated learning of eXplainable AI models. The Fed-XAI goal is achieved using post hoc explainability techniques [8], [9] or interpretable by-design models [10]–[13].

In this paper, we refer to FL of interpretable by-design models, focusing on a recent approach for FL of fuzzy rule-based classifiers (FRBCs) [14], which have proven effective in classification tasks on tabular data in heterogeneous setting, that is, when the local data of different participants follow different distributions. However, the distributed nature of FL expands the surface of attack that an adversary can exploit to interfere with the learning process of an ML model. In this work, our main focus lies on security threats related to adversarial attacks by malicious clients in FL of FRBC.

Security issues in traditional FL have been widely studied [15]–[17], typically assuming that FedAvg (or a variant) is used as the aggregation strategy and NN-based models are collaboratively learned. In particular, most attacks that a malicious adversary can implement are specific to such a class of models and aggregation strategy. For other approaches, it is crucial to understand how attacks can be crafted and to quantify the impact these may have on the learning process and the accuracy of the final model. In the Fed-XAI area, to the best of our knowledge, security issues have not been adequately investigated so far. The present work aims to partially fill this gap by analyzing the security threats to FRBC in FL. The main contributions of this work can be summarized as follows:

- we thoroughly describe the different types of attack a malicious client may implement and their impact on the FL of FRBC;
- we carry out a preliminary experimental analysis on two classification datasets to evaluate how the attacks affect the performance of the FRBC learned in a federated fashion;
- in the experimental analysis, the impact of the designed attacks for FL of FRBC is also compared with the impact of same or similar well-established attacks for traditional FL of NN-based models.

The rest of the paper is organized as follows. Section II describes background and related works on Fed-XAI and common attacks in traditional FL. Section III provides some preliminaries on FL of FRBC, while in Section IV we outline the security threats for the FL of FRBC. Section V describes the experimental setup and results on the impact of attacks on classification models. In Section VI, we draw conclusions.

II. Background and related Works

In this section, we first provide a brief background on FL and Fed-XAI, reviewing the most relevant recent work in the area. Then, we outline the security threats in traditional FL.

A. Federated Learning and Fed-XAI

Traditional FL [15], [18] typically involves a horizontal partitioning scenario (that is, clients have different samples over the same feature space) and a centralized communication topology, with the orchestration of a central server. Also, it deals mainly with the NN and DL models. The training stage of such models, in fact, is based on the optimization of a differentiable cost function, which can be easily achieved in the federated context by means of the iterative round-based FedAvg protocol [4]. In each round, the following steps are performed: (i) the central server sends the current global model to selected data owners; (ii) each selected data owner updates the model by performing some epochs of SGD on its local data; (iii) each selected data owner sends back the updated model to the server; (iv) the server takes the average of the locally updated models, weighted according to the number of samples, to obtain a new global model. Each local update shares the same structure (i.e., model architecture), making it readily possible to align models for aggregation.

Interpretable by-design models typically exploit optimization strategies that are not immediately compliant with FedAvg: as a consequence, their learning algorithm needs to be properly reworked to accommodate the FL setting. Several works in the Fed-XAI area have pursued this goal [6], [7]. The authors in [10] presented an approach for FL of a rule-based system for regression tasks. A federated version of the fuzzy C-Means algorithms is used to produce a global clustering of scattered data. Based on the discovered clusters, the parameters of the antecedent part of the rules are determined, whereas the parameters of the consequent part are adjusted with a federated gradient-based learning scheme. An alternative one-shot procedure has been proposed in [11]: Each data owner learns a rule-based model from its local data and sends it to the server. Then, the server aggregates the received rules by juxtaposing the rule bases collected from clients and by resolving possible conflicts. Federated versions of the decision tree induction algorithm have also been proposed. The authors in [13] proposed an approach for FL of a fuzzy regression tree: A single tree is generated on the server side, using aggregated statistics sent by the clients at each round.

Although FL for interpretable regression models has received some attention, only very few works have addressed their classification counterparts. The IBM FL framework [19] implements an adaptation of the ID3 algorithm for the FL setting. However, the framework is not open source.

Notably, all the above-mentioned works comply with the requirements of privacy and explainability, but none of them addresses the *security* aspect, which, however, is equally crucial to achieve trustworthiness in AI systems.

B. Security threats in traditional FL

Several recent surveys [15]–[17] provide comprehensive reviews of security in FL. In this section, we highlight the key elements of security threats in traditional FL, focusing on the typical scenario of horizontal FL and classification tasks, and using the taxonomy outlined in one of the most recent surveys [17].

In a centralized FL topology, insider attacks can be carried out by *clients* or by *server*. A further important distinction is made based on the attacker's goal: a *malicious* attacker tries to interfere with the FL process; an *honest-but-curios* attacker instead tries to obtain private information about other participants but still adhering to the FL protocol and without interfering with the learning process.

Attacks can be categorized according to several taxonomies. The attack objective distinguishes between *targeted or backdoor attacks*, aimed at injecting a secondary task into the model, and *untargeted attacks* aimed at damaging the model performance. In a backdoor attack, the adversary introduces an unwanted pattern into the system. The objective is to achieve high performance on the backdoor subtask without affecting the performance of the global model on its main task [20]. Consequently, the attack is particularly difficult to detect. A further taxonomy is proposed for training-time attacks from adversarial clients (which is one of the most common scenarios) and relates to the poisoned part of the FL process: data poisoning attacks and model poisoning attacks.

Data poisoning indicates a family of attacks in which the adversary manipulates the private data of corrupted clients. Traditional data poisoning attacks include random flipping, poisoning samples, and out-of-distribution attacks:

- in a *random flipping* attack the adversary alters a fraction of the labels associated with the training samples in the private local dataset by randomly reassigning them;
- in a *poisoning samples* attack the adversary modifies part of the training data samples, for instance, by injecting random noise;
- an *out-of-distribution* attack can be regarded as a special case of poisoning samples attack, in which poisoned samples are injected from outside the input distribution of local training data.

Model poisoning indicates a family of attacks in which the adversary directly manipulates the model parameters rather than corrupting the local data. By altering the model weights before sending them back to the server, the malicious user can cause the aggregated global model to behave incorrectly. Traditional data poisoning attacks include optimization methods and random weight:

- *optimization methods* are crafted to perform a backdoor attack, while minimizing the differences between the poisoned model and the aggregated model shared by the server at the last round;
- in a *random weights* attack the adversary randomly generates the local model updates of the corrupted client, i.e., the weights of the NN-based model. As a result, it affects the performance and reliability of the final global model.

In Section IV we present several possible data-poisoning and model-poisoning attacks that malicious clients can operate to FL of FRBC. For the preliminary experimental analysis, we will mainly focus on untargeted attacks in order to offer an unprecedented assessment of the impact of these attacks in the context of FL of explainable classifiers.

III. Preliminaries on Federated Learning of Fuzzy Rule-based Classifier

In this work, we analyze the impact of adversarial attacks against federated FRBCs. In the following, we first provide some preliminaries on FRBC and then present the approach for FL of FRBC considered in this paper, which was recently proposed in [14].

The knowledge base of an FRBC consists of a rule base (RB), composed of *if-then* rules, and a database (DB) containing the definition of the fuzzy sets used in the RB. The knowledge base is used to perform classification tasks, that is, classify any input instance into one of the K classes in a set $\Gamma = \{C_1, \ldots, C_K\}$. The generic m-th rule R_m of an RB is expressed as follows:

$$R_m : \textbf{IF } X_1 \textbf{ is } A_{1,j_{m,1}} \textbf{ AND} \ldots \textbf{AND } X_F \textbf{ is } A_{F,j_{m,F}} \\ \textbf{THEN } Y \textbf{ is } C_{j_m} with\ RW_m \quad (1)$$

where F is the total number of input features in the dataset, $A_{i,j_{m,i}}$ denotes the j^{th} fuzzy set of the fuzzy partition over the i^{th} input feature X_i, C_{j_m} is the class label associated with the rule, and RW_m is the rule weight. The latter term represents the degree of certainty of the classification in the class C_{j_m} within the subspace defined by the antecedent part of R_m, and is calculated based on the training samples that pertain to that region. Given a training set composed of N input–output pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, with $\mathbf{x}_t = [x_{t,1} \ldots, x_{t,F}] \in \mathbb{R}^F$, $t = 1, \ldots, N$ and y_t the associated label, the certainty factor CF_m is computed as the fraction between CF_NUM_m and CF_DEN_m:

$$RW_m = CF_m = \frac{\sum_{\mathbf{x_t} \in C_{j_m}} w_m(\mathbf{x_t})}{\sum_{t=1}^{N} w_m(\mathbf{x_t})} = \frac{CF\_NUM_m}{CF\_DEN_m} \quad (2)$$

Here, $w_m(\mathbf{x_t})$ is the activation strength, which quantifies the alignment between the t-th input instance and the antecedent of the generic m-th rule, and it is expressed by the formula:

$$w_m(\mathbf{x_t}) = \prod_{f=1}^{F} A_{f,j_{m,f}}(x_{t,f}) \quad (3)$$

In the inference stage, an RB with M rules can be used to determine the class of any given input instance $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \ldots \hat{x}_F\}$. Among possible reasoning methods, in this work we consider the *maximum matching* policy: the input instance is classified based on the rule with the highest association degree $h_m(\hat{\mathbf{x}})$, computed as:

$$h_m(\hat{\mathbf{x}}) = w_m(\hat{\mathbf{x}}) \cdot RW_m \quad (4)$$

The approach for FL of FRBC [14] is schematized in Fig. 1. It is designed to generate an FRBC in a collaborative and privacy-preserving way using a one-shot procedure, that is, a single communication round.

In the following, the FL process is described:

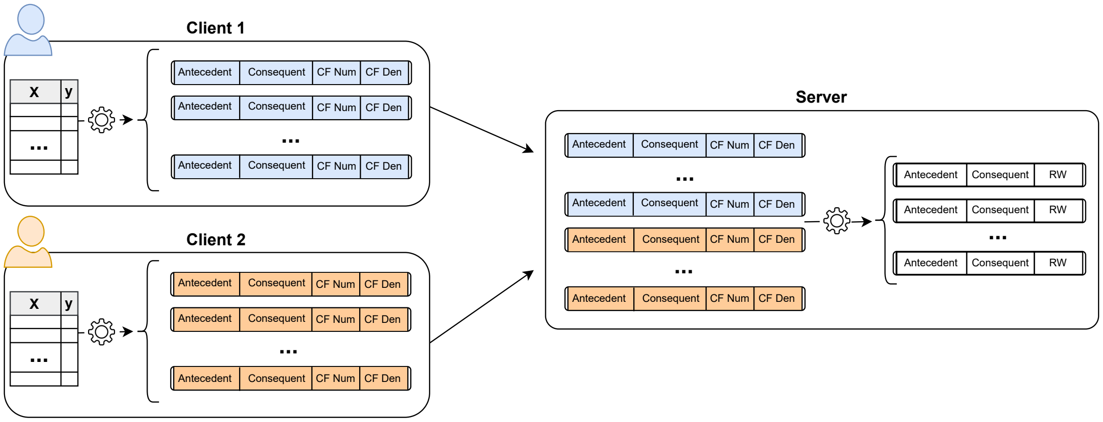


Fig. 1. Overview of the Federated FRBC algorithm [14], illustrated with a two-client toy example. Note that typically the FL process involves more clients.

- A centralized server configures the FL process by distributing to data owners a set of hyperparameters, defining the domain normalization and fuzzy partitioning for each input attribute.
- Each node uses its private data to independently generate a local RB. The rule generation phase uses the CHI algorithm [21]: a fuzzy partition is defined a priori on each input attribute; To improve interpretability, strong triangular fuzzy partitions are utilized [14]. Then, a rule is generated for each training sample. The antecedent of the rule is determined by selecting, for each attribute, the fuzzy set with the highest membership degree. Furthermore, the contributions CF_NUM_m and CF_DEN_m to, respectively, the numerator and the denominator in the formulation of CF (Eq. 2) is stored. Duplicate rules are discarded, and the final local RB is sent to the server for centralized processing. In particular, a local RB based on linguistic fuzzy rules does not reveal raw data, thus ensuring privacy preservation.
- The server aggregates the local RBs into a final RB. Rules without duplicates or conflicts are directly inserted into the final model. For each set DR_m of duplicate rules (i.e., same antecedent and same consequent), a single rule R_m is included in the final RB and the weight RW_m is calculated as the ratio of the aggregated numerator to the aggregated denominator contributions:

$$RW_m = \frac{\sum_{R_m^i \in DR_m} Num_m^i}{\sum_{R_m^i \in DR_m} Den_m^i}. \tag{5}$$

  In the case of conflicting rules (that is, same antecedent and different consequent), only the rule with the highest weight is included in the final RB.

The generated FRBC represents the federated model and can be distributed to each node for inference purposes.

IV. Security Threats to FRBC in FL

The procedure for FL of interpretable FRBC described in Section III naturally deviates from the traditional FL approaches, which typically relies on FedAvg. In FedAvg, model updates are shared with the server at each round, whereas in FL of FRBC, clients share rule-based models derived from private data for centralized one-shot aggregation.

In this section, we illustrate possible attacks that a malicious client can implement within the context of FL of FRBC. To this end, we consider the following setup: The malicious client has access to local data and the model, and it has complete knowledge of how the entire FL algorithm works, including both the local learning phase and the global aggregation phase. Notably, the server is not directly affected by any adversary: we assume a *honest-but-curious* server, as is typical in the horizontal FL literature [5], which always adheres to the protocol defined for the execution of the ML algorithm.

In the following, we outline the adversarial attacks to FL of FRBC distinguishing between the targeted and untargeted ones. The diagrams in this section are derived from Fig. 1, assuming that *Client 1* is malicious.

A. Targeted attack: Backdoor

A schematic representation of a targeted attack (backdoor) to the FL of FRBC is shown in Fig. 2.

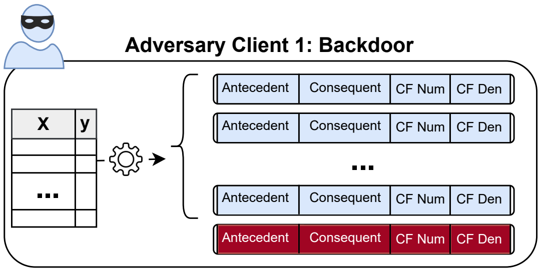


Fig. 2. Backdoor attack. The adversary injects the desired rule directly in the local RB of the corrupted client.

The global interpretability of the FRBC enables the adversary to immediately detect which rules are generated in the local RB. Thus, a backdoor is particularly easy to create in this context. The malicious client can inject a new rule into the local RB to accomplish a desired classification task as follows: the antecedent part identifies a specific fuzzy region in the feature space (typically different from those captured by existing rules), the consequent part holds the desired classification label, and the contribution to rule weight is arbitrarily set to promote the rule in case of conflicts in the aggregation stage.

The backdoor attack typically has little or no impact on performance. In fact, the influence of a single rule is localized in the feature space and only comes into play when it is prioritized during conflict resolution or selected at inference time based on the maximum matching policy. For this reason, the experimental analysis proposed in this paper focuses on untargeted attacks, which damage the performance of the overall model, as detailed in the next section.

B. Untargeted attacks

Untargeted attacks aim to damage the performance of the global model. Based on the literature review reported in Section II, we envision five untargeted attacks to FRBC in FL: random flipping, out-of-distribution, random weights, random consequent, random rule, schematically represented in Fig. 3.

Random flipping (Fig. 3a) implements the data poisoning attack, as in traditional FL: it consists of the manipulation of target labels in a subset of the private training set of a corrupted client. The attack affects two key aspects of the resulting RB. First, the rules generated from the altered data samples will reflect the modified class label. Second, the contribution to the numerator of the rule weight is altered for all rules activated by the modified samples.

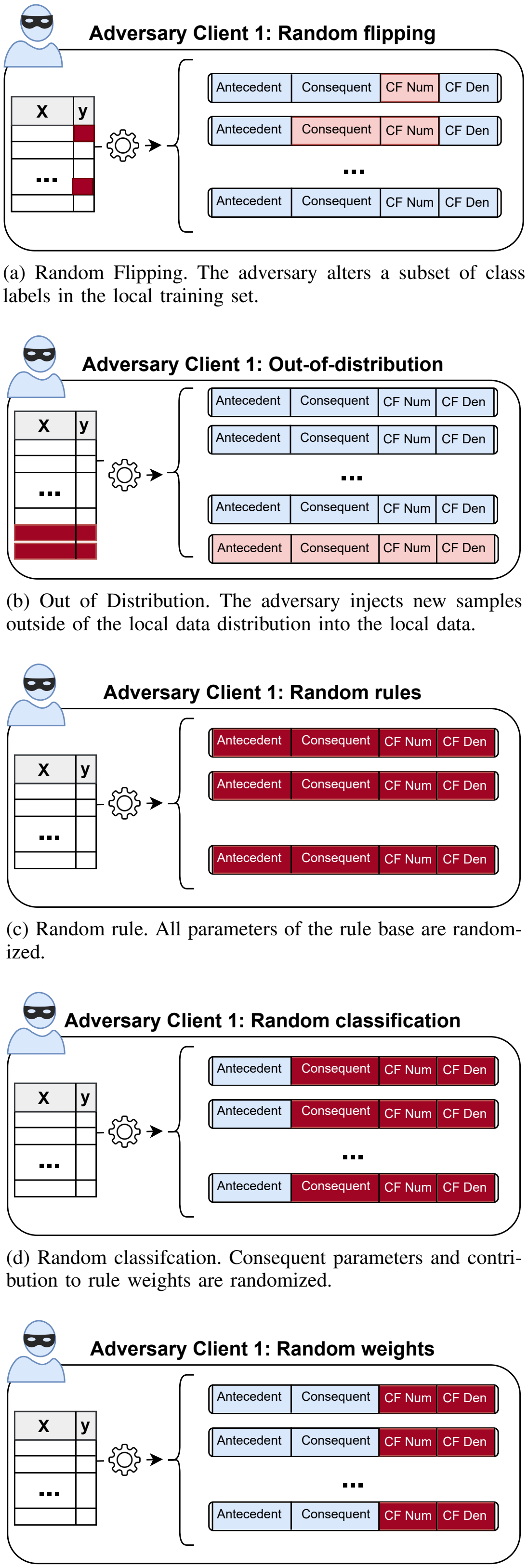
The *out-of-distribution* (Fig. 3b) data poisoning attack introduces new samples into the private training set of a corrupted client, as in traditional FL. However, unlike traditional FL, this type of attack has an effect that closely resembles backdoor attacks in the context of FRBC. In fact, the CHI algorithm used for local rule generation tends to create a new rule for each injected sample, which, by definition, is crafted to deviate from the distribution of the client's original data. As a consequence, new rules are included in the local RB, with no or limited effect on the rules generated by the in-distribution samples.

While the above-mentioned data poisoning attacks are obviously agnostic with respect to the model learned in a federated fashion, model poisoning attacks need to be revisited in the context of FRBC. They can be implemented by altering specific subsets of the rules components.

A *random rule* (Fig. 3c) attack involves randomizing all components of the rules. It is the most disruptive model poisoning attack, meaning that it completely overrides the rule-based generated by local data.

A *random classification* (Fig. 3d) attack randomizes all but the antecedent part of the rules. By altering the consequent part and the contribution to the rule weights, it potentially introduces new rules into the final model and impacts the conflict resolution stage.

A *random weight* (Fig. 3e) attack modifies only the contribution to the rule weights (local statistics for the computation of the certainty factor). As a consequence, it does not introduce new rules, but only impacts the conflict resolution stage.



Adversary Client 1: Random flipping
X
y
...
Antecedent
Consequent
CF Num
CF Den
Adversary Client 1: Out-of-distribution
Adversary Client 1: Random rules
Adversary Client 1: Random classification
Adversary Client 1: Random weights

(a) Random Flipping. The adversary alters a subset of class labels in the local training set.

(b) Out of Distribution. The adversary injects new samples outside of the local data distribution into the local data.

(c) Random rule. All parameters of the rule base are randomized.

(d) Random classifcation. Consequent parameters and contribution to rule weights are randomized.

(e) Random weights: Contribution to rule weights are randomized.

Fig. 3. Security attacks to FRBC in FL. Dark red indicates the poisoned part of the attack; Light red indicates a possible indirect modification on the locally learned rule-based system.

V. Experimental analysis

This section presents a preliminary experimental analysis to assess the impact of the threats described in Section IV on FRBC in FL. We first describe the experimental setup and then discuss the experimental results. Also, we report on a comparative analysis between FL of FRBC and traditional FL of NN-based model, evaluating the impact of same or similar well-established attacks. For the NN-based model, we consider a multilayer perceptron (MLP).

A. Experimental setup

In our experiments, two binary classification datasets are considered as in [14]: MAGIC Gamma Telescope [22] and Quality of Experience (QoE) for video streaming [23]. Data among clients are forced to be non-i.i.d. (non-independent and identically distributed), as is typical in horizontal FL, considering 20 clients for Magic and 30 clients for QoE.

We implemented the proposed *untargeted attacks* to FL of FRBC for both case studies, with an increasing percentage K of corrupted clients. We consider the case $K = 0\%$ as our baseline, i.e., no malicious clients, and evaluate each attack for $K \in \{5\%, 15\%, 25\%, 35\%\}$. Each configuration is repeated three times with different random seeds to sample the clients that act maliciously. We will report the average values in the results. To assess the performance of the federated FRBC across various configurations, we resort to centralized test sets, which follow the overall data distribution for both datasets.

Table I reports summary statistics of the adopted datasets.

TABLE I
Datasets description

	Num. of features	Training set size (overall)	Num. of clients	Num. of malicious clients	Test set size
Magic	10	15210	20	$\{0, 1, 3, 5, 7\}$	3810
QoE	30	25311	30	$\{0, 2, 4, 7, 10\}$	6505

In the following, we report the parameters configuration for FL of FRBC, for FL of MLP and for the attacks.

FL of FRBC requires setting the number T_f of fuzzy sets used to partition each input attribute $f \in F$. Following guidance from the relevant literature [14], we set $T_f = 5$ for Magic and $T_f = 3$ for QoE, for each feature, as these values have been proven to ensure an optimal balance between performance and explainability. The MLP architecture consists of two fully connected layers with ReLU activation functions, containing 256 and 128 neurons, respectively. The output layer uses a sigmoid activation function to classify sample instances into two possible classes. The number of local epochs is set to 3, the minibatch size is set to 32, and the number of rounds is set to 10. We did not perform extensive hyperparameter optimization; instead, we simply ensured that the classification results align with the state-of-the-art and focused primarily on analyzing the impact of security threats. Notably, all attacks described in the following are applied independently to each malicious client.

Data poisoning attacks are carried out identically for both models. In random flipping (RF) we randomly sample a label in the set $\Gamma = \{0, 1\}$ for each training sample. We recall that both case studies involve binary classification tasks and therefore, on average, half of the labels are altered. The out-of-distribution (OoD) attack is executed by injecting samples equal to 5% of the size of the training set. The generation of OoD instances involves a uniform random sampling of values within the range of each input feature. Each generated sample is assigned a randomly sampled label from Γ.

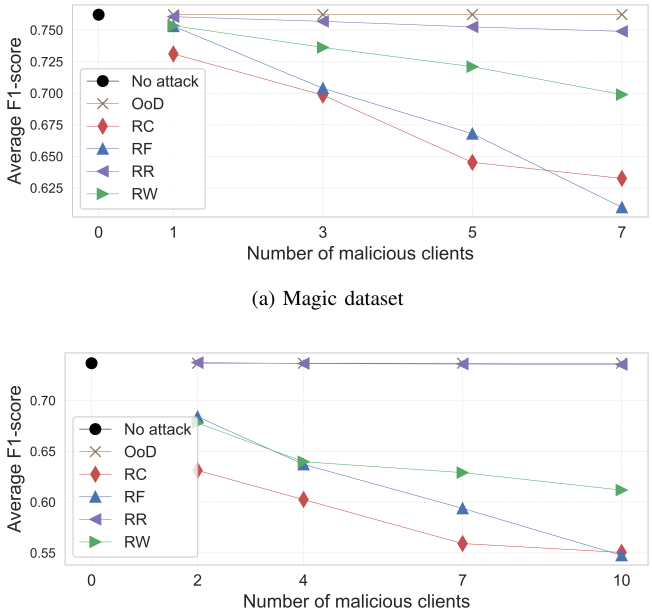
For FRBC, model poisoning attacks are configured as follows. In random weights (RW), the contribution to the numerator and to the denominator in the formulation of CF is randomly sampled from a uniform distribution in $[0, 100]$. Random classification (RC) extends RW by also applying a perturbation to the consequent part of the rules, that is, randomly sampling the label from Γ. Finally, the random rule (RR) involves the same steps as the RC and also alters the antecedent part of the rules by randomly sampling a fuzzy set in the fuzzy partition of each attribute.

For MLP, RW is obtained by generating the model updates with uniform random sampling of values within the range $[-0.5; 0.5]$.

Performance results are reported as the macro-average F1 score on the test set. For FRBC, we also report the number of rules of the global model. This highlights the impact of attacks that introduce new patterns into the final system, providing further insight into the consequences of adversarial behavior.

B. Experimental results: attacks to FL of FRBC

Figure 4 illustrates the impact of adversarial attacks on federated FRBC in terms of the average F1-score on both datasets, as the number of corrupted clients increases.



(a) Magic dataset

(b) QoE dataset

Fig. 4. Impact of attacks on FL of FRBC in terms of average F1-Score, with increasing number of malicious clients.

In general, the extent of performance degradation depends on the type of attack. As expected, the drop in F1-score tends to worsen as the number of malicious clients increases. For the most impactful attack, performance severely deteriorates when 35% of the clients are malicious: F1-score drops from 0.76 to 0.61 (around 20%) and from 0.74 to 0.55 (around 25%) in the Magic and QoE datasets, respectively.

In the Magic data set (Fig. 4a), the attacks that most significantly degrade the performance of the model are RF and RC. In these cases, the antecedent parts of the rules remain unchanged, and thus they model regions of the feature space are actually relevant to the dataset. Instead, the consequent part (and the class label in particular) is randomized, significantly undermining the model classification ability. The OoD and RR attacks have minimal impact on performance: they tend to introduce additional rules into the final FRBC, but such rules, being randomly generated, do not model the distribution of real data. As a consequence, they are seldom activated during inference on the test. The RW attack, in which only the contributions to rule weights are altered, still shows a noticeable effect on the performance of the model: rule weights, in fact, play a key role for the conflict resolution stage and for rule selection at inference time.

Similar results are observed for the QoE dataset (Fig. 4b): RF and RC attacks are the most harmful, whereas RW has a milder effect. RR and OoD attacks do not affect model performance at all.

In general, we observed that adversarial attacks can potentially modify existing rules and create new ones, affecting both the metrics and the complexity of the global FRBC. Table II reports the impact of the various attacks on the number of federated FRBC rules for the two datasets, with an increasing number of corrupt clients.

TABLE II
IMPACT OF ATTACKS ON THE NUMBER OF RULES OF THE FEDERATED FRBC, WITH INCREASING NUMBER OF MALICIOUS CLIENTS. THE NUMBER OF RULES IN CASE OF NO ATTACK IS REPORTED WITHIN PARENTHESIS FOR EACH DATASET. AVERAGE VALUES.

	Percentage of malicious clients			
Attack	5%	15%	25%	35%
Magic (1800)				
RF, RC, RW	1800	1800	1800	1800
OoD	1821	1893	1931	1981
RR	2018	2542	2960	3491
QoE (4809)				
RF, RC, RW	4809	4809	4809	4809
OoD	4883	4951	5052	5175
RR	5559	6232	7190	8437

RR attack clearly has the most significant impact, as malicious client injects new random rules into the system. Even in the case of OoD attacks, some additional rules are introduced, albeit to a lesser extent compared to RR. In fact, it is worth recalling that the OoD samples injected for each client amount to only 5% of the size of the local training set. Obviously, the other attacks (RF, RC, RW) do not alter the number of rules, since none of them modifies the antecedent part of the rules.

C. Comparative analysis: attacks to FL of FRBC and to FL of MLP

MLP slightly outperforms FRBC in both datasets. The average F1-score is 0.78 on Magic, and 0.76 on QoE. To assess the relative impact of attacks on the two models, we report the percentage degradation of the average F1-score in Fig. 5, with an increasing number of malicious clients.

The FRBC is much more susceptible than the MLP to the RF attack on both datasets: on the one hand, in the rule-based system, noisy labels are directly reflected in conflicting or incorrect rules in the global model. On the other hand, we argue that the MLP is more robust to such noise because the local adjustment of weights and the global averaging aggregation stage tend to mitigate its impact. For OoD attacks, both models show minimal degradation.

The effect of RW attacks is considerable for both models. As we have discussed in the previous section, for FL of FRBC the impact on performance depends on whether model poisoning involves randomizing the contribution to rule weights and consequents (RC, highest impact), only the contribution to rule weights (RW, medium impact) or all the rule parameters (RR, lowest impact). For FL of MLP, the RW attack consists of randomizing the values of the weights: The effect is limited on the QoE dataset while there is a noticeable performance degradation on the Magic dataset with a high number of malicious clients.

In general, the FRBC is less robust than the MLP to the RF data-poisoning attack and also exhibits considerable vulnerability to certain model-poisoning attacks (RC and RW). Defense mechanisms aimed at enhancing the security of FL of FRBC should be oriented towards identifying these attacks and mitigating their effects. Refinement of rule aggregation processes and integration of advanced adaptive weighting and conflict resolution strategies could significantly enhance the robustness of FRBC while preserving their inherent interpretability. In addition, the interpretable nature of FRBCs could be leveraged to evaluate candidate rules shared by clients, enabling identification of potential backdoor attacks.

VI. CONCLUSION

In this paper, we investigated the impact of adversarial attacks to explainable models, namely fuzzy rule-based classifiers (FRBCs), in federated learning (FL). First, we present the attacks that malicious clients can operate to federated FRBC, considering the peculiarities of the model and the aggregation strategy, necessarily different from those of the traditional FL setting. Second, an experimental analysis was performed to assess the robustness of federated FRBC under untargeted attacks (i.e., those aimed at damaging the model performance) by using two real-world case studies. Third, we compared the effect of same or similar attacks to federated FRBC and to federated black-box model, namely Multi-Layer Perceptron. The results revealed that the FRBC exhibits high susceptibility to untargeted attacks, particularly those involving label manipulations or alterations to the local rule bases. Future work will focus on the defense mechanism against adversarial attacks to

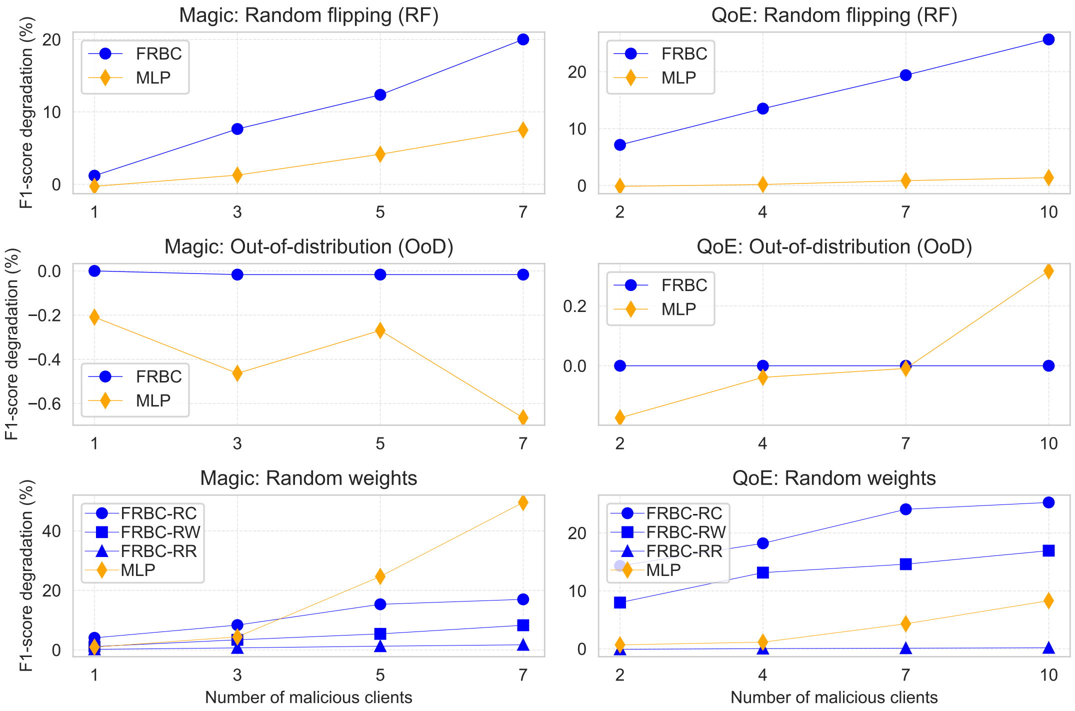


Fig. 5. Comparison of the impact of adversarial attacks on FL of FRBC and on FL of MLP in terms of percentage degradation of average F1-score.

federated FRBC, aiming to enhance security while maintaining the inherent privacy and explainability of the AI system, which are considered pivotal for its trustworthiness.

References

[1] High Level Expert Group on AI, "Ethics Guidelines for Trustworthy AI, Technical Report," 2019, EC. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inform. Fusion*, vol. 58, pp. 82–115, 2020.

[3] R. Guidotti, A. Monreale *et al.*, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018.

[4] B. McMahan, E. Moore *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics*, ser. Proc. of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[5] Q. Yang, Y. Liu *et al.*, "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, jan 2019.

[6] J. L. Corcuera Bárcena, M. Daole *et al.*, "Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models," in *XAI.it: 3rd Italian WS on Explainable Artificial Intelligence, co-located with AI*IA*, 2022.

[7] R. López-Blanco, R. S. Alonso *et al.*, "Federated Learning of Explainable Artificial Intelligence (FED-XAI): A Review," in *Distributed Computing and Artificial Intelligence, 20th Int'l Conf.* Cham: Springer Nature Switzerland, 2023, pp. 318–326.

[8] J. Fiosina, "Interpretable Privacy-Preserving Collaborative Deep Learning for Taxi Trip Duration Forecasting," in *Int'l Conf. on Vehicle Technology and Intelligent Transport Systems, Int'l Conf. on Smart Cities and Green ICT Systems*. Springer, 2022, pp. 392–411.

[9] A. Bogdanova, A. Imakura, and T. Sakurai, "DC-SHAP Method for Consistent Explainability in Privacy-Preserving Distributed Machine Learning," *Human-Centric Intelligent Systems*, vol. 3, no. 3, pp. 197–210, 2023.

[10] X. Zhu, D. Wang *et al.*, "Horizontal Federated Learning of Takagi–Sugeno Fuzzy Rule-Based Models," *IEEE T FUZZY SYST*, vol. 30, no. 9, pp. 3537–3547, 2022.

[11] J. L. Corcuera Bárcena, P. Ducange *et al.*, "An Approach to Federated Learning of Explainable Fuzzy Regression Models," in *2022 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2022, pp. 1–8.

[12] L. J. Dust, M. L. Murcia *et al.*, "Federated Fuzzy Learning with Imbalanced Data," in *2021 20th IEEE Int'l Conf. on Machine Learning and Applications (ICMLA)*, 2021, pp. 1130–1137.

[13] J. L. Corcuera Bárcena, P. Ducange *et al.*, "Increasing trust in AI through privacy preservation and model explainability: Federated Learning of Fuzzy Regression Trees," *Inform. Fusion*, vol. 113, p. 102598, 2025.

[14] M. Daole, P. Ducange *et al.*, "Trustworthy AI in Heterogeneous Settings: Federated Learning of Explainable Classifiers," in *2024 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2024, pp. 1–9.

[15] P. Kairouz, H. B. McMahan *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[16] V. Mothukuri, R. M. Parizi *et al.*, "A survey on security and privacy of federated learning," *Future Gener. Comp. Sy.*, vol. 115, pp. 619–640, 2021.

[17] N. Rodríguez-Barroso, D. Jiménez-López *et al.*, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Inform. Fusion*, vol. 90, pp. 148–173, 2023.

[18] B. McMahan, E. Moore *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics*, ser. Proc. of Machine Learning Research, vol. 54. PMLR, 2017, pp. 1273–1282.

[19] H. Ludwig, N. Baracaldo *et al.*, "IBM Federated Learning: an Enterprise Framework White Paper V0.1," 2020.

[20] E. Bagdasaryan, A. Veit *et al.*, "How To Backdoor Federated Learning," in *Proc. of the 23 Int'l Conf. on Artificial Intelligence and Statistics*, vol. 108. PMLR, 26–28 Aug 2020, pp. 2938–2948.

[21] Z. Chi and H. Yan, *Fuzzy algorithms: with applications to image processing and pattern recognition*. World scientific, 1996, vol. 10.

[22] R. Bock, "MAGIC Gamma Telescope," UCI Machine Learning Repository, 2007, DOI: https://doi.org/10.24432/C52C8B.

[23] J. L. C. Bárcena, P. Ducange *et al.*, "Towards Trustworthy AI for QoE prediction in B5G/6G Networks," in *First Int'l Workshop on Artificial Intelligence in Beyond 5G and 6G Wireless Networks (AI6G)*, 2022.