Rule-Based Federated Learning for Healthcare

Kamil Wozniak[1][0000−0002−2342−3407] ✉, Jose Sousa[1,2][0000−0001−9570−6054], and Bartlomiej Sniezynski[3][0000−0002−4206−9052]

[1] Sano – Centre for Computational Personalised Medicine, Krakow, Poland https://sano.science/
[2] Multidisciplinary Institute of Ageing, MIA-Portugal, University of Coimbra, Coimbra, Portugal https://www.uc.pt/mia/
[3] AGH University of Krakow, Krakow, Poland https://www.agh.edu.pl/
k.wozniak@sanoscience.org

Abstract. This paper investigates the adaptation of classical rule-based classification algorithms (LEM2, PRISM, and RIPPER) to the federated learning paradigm. We propose a federation discretization framework that establishes global bin boundaries using aggregated feature statistics, ensuring consistent rule vocabulary across clients without sharing raw data. We conducted experiments on five medical datasets, which had varying numbers of clients (2–5) and discretization granularities (2–7 bins), with all configurations evaluated across 5 random data splits. The results show that the global model consistently achieves high levels of predictive accuracy and interpretability, outperforming averaged results of local models by 7.0-9.1% in balanced accuracy across algorithms. Notably, federated LEM2 significantly outperformed even its centralized counterpart ($p = 0.011$), while federated RIPPER and PRISM achieved statistical parity with centralized baselines ($p > 0.05$). On average, the resulting rulesets are compact (8.3–53.1 rules) and interpretable (1.39–4.26 conditions per rule), with coverage ranging from 51% to 96% depending on the algorithm. These results demonstrate that federated rule learning can provide both transparency and competitive performance for clinical applications where data is limited and privacy is essential.

Keywords: Biomedical Data · Federated Learning · Rule Induction · Interpretability.

1 Introduction

Tabular data forms the backbone of many real-world artificial intelligence (AI) and machine learning (ML) applications [8]. This is particularly evident in domains such as finance, healthcare, commerce, and public administration [33]. In healthcare, for example, data in this format is commonly encountered, covering aspects such as laboratory test results, patient demographics, prescribed medications and dosages, treatment responses, applied therapies, and ultimately diagnoses themselves.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

An important characteristic of medical data is its high sensitivity, necessitating the prioritization of security and patient privacy [26]. This is often accompanied by complex regulations such as General Data Protection Regulation (GDPR), which complicate working with such data and frequently prevent effective centralization of data from different institutions (referred to as clients). This leads to the emergence of a need for effective data utilization without compromising security or risking exposure to unauthorized parties presenting a challenge that has largely driven the development of federated learning (FL) [36, 19]. Furthermore, many healthcare applications require models to be sufficiently explainable to allow clinicians to understand why a particular prediction was made [31].

FL enables independent entities to collaborate in training ML models without sharing raw data [16]. Neural networks currently dominate this area due to the natural compatibility of their gradient-based model optimisation with collecting updates from many different sources [23]. Despite offering high scalability, neural networks, as opaque models, suffer from limited transparency, which is a critical concern in highly regulated domains [3, 28].

At the same time, classical machine learning methods such as decision trees or rule mining algorithms were not designed with efficient learning in distributed, privacy-preserving settings in mind. These methods, however, often achieve high performance on tabular data, especially when data volume is limited, achieving competitive performance while exhibiting a much higher degree of interpretability [13]. They stand in contrast to black-box models whose internal mechanisms are difficult to understand.

Models sometimes referred to as good old-fashioned artificial intelligence (GOFAI) do not lie at the center of interest of researchers working on FL, and work in this area remains limited [9, 19]. A natural question related to rule-based methods in the context of FL is whether these models can be successfully utilized in medical settings using FL while maintaining high prediction accuracy and a high level of explainability. A secondary consideration here is also the general FL concern of the privacy versus accuracy trade-off and the overall costs associated with communication between different institutions in real-world applications.

The contributions of this work include the design, implementation and evaluation of three federated rule-based algorithms (LEM2, PRISM, RIPPER) for distributed medical data; the demonstration that federated rule learning achieves high prediction quality while preserving full interpretability; a comprehensive experimental evaluation across five biomedical datasets with varying numbers of clients and discretization settings; and the validation of the value of federated collaboration.

2 Background

FL is a distributed ML framework that prioritizes data privacy through decentralization [35]. Google introduced it in 2016 to enable on-device training of deep neural networks [23]. At the core of this paradigm is the assumption that local

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

raw data is not shared. Fundamental to it is the separation between a central coordinator (server) and institutions or devices (clients).

The most general division of FL into different types is made based on how data is partitioned among individual clients. The fundamental type here is horizontal FL, in which all clients share the same feature space while having disjoint samples of the data (e.g., hospitals may collect blood marker measurements from different patients). The horizontal approach is also the focus in this work.

The horizontal model is trained collaboratively by clients under the supervision of the server. In the standard approach, the goal is to minimise the global cost function, for which there exist various federated optimisation algorithms. The most popular is Federated Averaging (FedAvg), which involves averaging local model weights to build a global model [23].

Standard FL algorithms such as FedAvg aggregate model weight updates across clients [23]. Key FL challenges include not independent and identically distributed (non-IID) data across clients, communication overhead, and heterogeneous client resources [19, 21]. Furthermore, rule-based methods require alternative aggregation strategies, as discussed in Methods section.

The choice of model type within the FL framework has a fundamental impact on both the predictive power of the model and the transparency of its decisions [19]. While neural networks dominate FL, due to the high compatibility of gradient-based optimization with this framework, their black-box nature limits their transparency, which is a growing concern, reflected in regulatory requirements such as the EU's GDPR right to explanation [12, 15].

Rule-based algorithms are representatives of the symbolic machine learning family, where dependencies extracted from data are expressed in the form of human-readable decision rules [9, 24]. Each rule of this type has the form: IF conditions THEN class, where conditions are a conjunction of one or more attribute-value pair conditions (e.g., IF CRP = high AND smoking = yes THEN diagnosis = sick). In contrast to neural networks, each decision made by such a model can be explained by referring to a specific rule.

This characteristic makes these methods particularly suitable for critical and highest-risk applications, where accountability and transparency are crucial. Classical approaches to rule induction include, among others, covering algorithms [10], which iteratively learn rules one by one in order to cover training examples, as well as methods based on rough set theory, which induce rules from data reducts [25].

3 Methods

3.1 Global Discretization Strategy

Rule-based algorithms naturally expect discretized features and clients share consistent bin boundaries to make rules comparable across clients. In the proposed approach, each client c for each feature j computes local minimum and maximum values:

$$v_{\min,j}^{(c)} = \min(X_j^{(c)}), \qquad v_{\max,j}^{(c)} = \max(X_j^{(c)}) \tag{1}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

where $X_j^{(c)}$ denotes the values of feature j in client c's local dataset.

Clients send the computed local statistics to the central server. The server collects this information and computes global minimum and maximum values:

$$v_{\min,j}^{\text{global}} = \min_c v_{\min,j}^{(c)}, \qquad v_{\max,j}^{\text{global}} = \max_c v_{\max,j}^{(c)}, \qquad j = 1, \ldots, p \tag{2}$$

The server broadcasts these global values to all clients. Clients then perform equal-width discretization with a specified number of bins k, using bin boundaries:

$$b_{j,i} = v_{\min,j}^{\text{global}} + i \cdot \frac{v_{\max,j}^{\text{global}} - v_{\min,j}^{\text{global}}}{k}, \quad i \in \{0, 1, \ldots, k\} \tag{3}$$

In this way, all clients operate on identical bins, ensuring that locally induced rules are directly comparable and can be meaningfully aggregated.

3.2 Federated Rule Learning Algorithms

We designed FL versions of three rule-based algorithms: LEM2, PRISM and RIPPER. All three naturally operate on discretized data, which makes them compatible with the global discretization strategy described above. Each algorithm builds local rulesets, which are then aggregated on the server using the aggregation strategy described below. Below, we describe the characteristics of each algorithm.

Federated LEM2 LEM2 (Learning from Examples Module, version 2) is an algorithm for rule induction based on rough set theory [14]. The algorithm operates on the principle of maximal generalization, seeking the most general rules that correctly classify training examples, helping to prevent overfitting on fragmented local data. In the federated setting, each client builds a local LEM2 model using the global bin boundaries received from the server. Rules are induced based on the local training set, with conditions expressed as conjunction of discretized attribute-bin pairs. LEM2's maximal generalization principle tends to produce numerous rules with relatively few conditions each, resulting in broad coverage of the feature space.

Federated PRISM PRISM is a rule induction algorithm based on the separate-and-conquer strategy [5]. For each class, it iteratively searches for a conjunction attribute-value pairs that maximize the conditional class probability, a process requiring an appropriate number of examples per class to be reliable. Each client extracts rules from local training data using the global bin boundaries and sends them to the server for aggregation. PRISM tends to produce fewer but more complex rules than LEM2.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

Federated RIPPER RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is a sequential covering algorithm optimized for extracting concise rulesets [11, 6]. It implements a three-stage process: growing, pruning (requiring a sufficient amount of local validation data to prevent overfitting) and optimization. Each client trains a local model and sends the resulting ruleset to the server. The main advantage of RIPPER emerges from its aggressive optimization: it creates effective and concise rulesets, usually much more compact than those of PRISM and especially LEM2.

3.3 Aggregation Strategy

After local rules extraction, they must be combined in order to create a global model. For this purpose, in this work we employ union aggregation, which gathers all unique rules from all clients (two rules are considered identical if they have the same set of attribute-value pairs as conditions):

$$R_{\text{global}} = \bigcup_{c=1}^{C} R_c \tag{4}$$

where R_c denotes the ruleset from client c and C is the total number of clients. This strategy allows preserving each locally detected complementary patterns in the data, thus ensuring maximum diversity and coverage of the global model.

For algorithms that produce ordered rules, the local ordering within each client's ruleset is preserved, but no global re-ordering is performed. As a result, during inference, for rules originating from the same client, more specific rules (with more conditions) are checked before more general ones. However, since rulesets are applied sequentially, a more general rule from an earlier client may be applied before a more specific rule from a later client.

3.4 Inference

Prediction is made based on the same global bin boundaries that were used during local rule learning. This ensures consistency between training and inference. Given a data instance to be predictively classified, its features are first discretized using global bins and then matched against rules in the globally aggregated ruleset.

The first rule whose conditions are satisfied determines the class predicted. A condition in the form of feature-bin is satisfied when it matches the feature-bin value of the data instance. The order of rules in the global ruleset reflects the order in which the results of individual clients are received. The first-match approach additionally ensures computational efficiency of the prediction. Furthermore, if no rule is matched, a default rule is invoked which assigns the default class of the considered instance. The frequency with which this default rule is applied is reflected in the value of the $1 - \textit{coverage}$ metric presented in the results.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

4 Experimets

4.1 Datasets

Five medical datasets were used: Breast Cancer Wisconsin (569 instances, 30 features) [34], Cleveland Heart Disease (297, 13) [18], Chronic Kidney Disease (195, 22) [27], Parkinsons (195, 22) [22] and Pima Indians Diabetes (768, 8) [29].

4.2 Experimental Setup

To ensure statistical robustness and mitigate the effect of data partitioning, all experiments were conducted five times using different random seeds. For each seed, the dataset was randomly split into training (80%) and test (20%) sets. The training set was equally divided among C clients. The test set was retained as a global test set for model evaluation. Experiments were conducted across varying number of clients ($C \in \{2, 3, 4, 5\}$) and discretization bins ($k \in \{2, 3, 4, 5, 6, 7\}$) using federated LEM2, PRISM and RIPPER on 5 datasets (see Section 4.1). This resulted in a total number of $4 \times 6 \times 3 \times 5 \times 5 = 1800$ experiments. The default values of available hyperparameters for all algorithms were used, and all experiments were implemented in Python using the Flower framework for FL simulations [1].

4.3 Metrics

Balanced accuracy was selected as the primary metric due to class imbalance in most datasets [4]. Additionally, the statistical significance of differences in prediction quality between different versions of the models was assessed using paired Wilcoxon signed-rank test with a significance threshold of $\alpha = 0.05$ [7]. The tests were conducted using performance scores averaged across all configurations for each of the 5 random seed splits for each dataset.

To quantify the degree of interpretability of the created global rulesets, we measured: ruleset size (number of rules in the global ruleset), rule complexity (average number of conditions per rule) and coverage (proportion of test cases that matched a rule other than the default rule).

Additionally, to assess the diversity of rules obtained by clients, we calculated similarity between individual local rulesets. Rules in the form of tuples of sorted conditions are treated as elements of sets, the similarity of which is determined using Jaccard similarity value given by the quotient of the number of shared rules and the total number of unique rules across both rulesets i and j [30]:

$$J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \tag{5}$$

This value was averaged across all client pairs (i, j). The smaller the value of this metric, the more diverse and complementary the rules are between two clients. A low value thus indicates that global aggregation allows capturing a broader range of patterns present in the data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

In order to investigate and quantify the gain resulting from the application of federated aggregation, we compared the balanced accuracy values of global models with averaged values of local models. Each local model was tested separately on the global test set, and the performance of individual models was averaged. Furthermore, the performance of global federated models was compared against their fully centralized counterparts to evaluate the degree of prediction quality degradation resulting from data distribution across local clients.

5 Results

We performed a series of experiments using three FL-adapted algorithms (LEM2, PRISM, RIPPER) and 5 datasets with varying configurations. The total number of experiments amounted to 1800, during which the number of bins (from 2 to 7) and the number of clients (from 2 to 5) were systematically changed for 5 different data splits. Moreover, in each experiment, the data were discretized using equal-width binning and the union aggregation strategy was employed.

The main research question was whether rule-based federated models can be effectively trained in a federated way on small biomedical data while preserving their high prediction quality and degree of interpretability. Our results show that federated rule learning is viable, achieving competitive results relative to their centralized counterparts and additionally consistently outperforming averaged results of local models (improvement of 7.0 % to 9.1 % in balanced accuracy across algorithms). Furthermore, statistical tests showed that federated LEM2 significantly outperformed even the model trained in a centralized way ($p = 0.011$), while federated RIPPER and PRISM achieved statistical parity with centralized baseline models ($p > 0.05$). These results suggest that global aggregation of rules leads to capturing more broad patterns in the data which in turn results in better model generalization ability compared to separate local models. Figure 1 summarizes the balanced accuracy, revealing the average response of algorithms to changes in the number of clients and bins. It can be seen that federated LEM2 and RIPPER exhibits similar, noticeably better than federated PRISM, stability across configurations.

5.1 Algorithms Performance Comparison

Table 1 presents the overall performance of the three federated algorithms, averaged across data splits and configurations. Federated LEM2 achieved the highest overall average balanced accuracy of 0.82 (SD = 0.13), followed closely by federated RIPPER with 0.79 (SD = 0.14). Federated PRISM demonstrated notably lower performance with a mean of 0.66 (SD = 0.15). The lowest standard deviation for federated LEM2 suggests lower variability in performance across different datasets and configurations, indicating potentially higher robustness compared to PRIM and RIPPER.

Table 1 also presents the best achieved balanced accuracy for each algorithm across the five datasets (Best sub-column). The results reveal dataset-specific patterns: All three algorithms achieved markedly high performance on

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

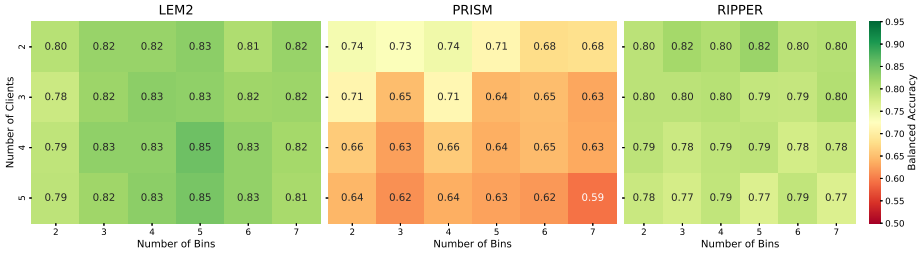

Fig. 1. Heatmaps showing balanced accuracy as a function of the number of discretization bins (columns) and the number of clients (rows) for each algorithm, averaged across all 5 datasets and 5 random data splits. Darker green indicates higher performance.

Table 1. Federated algorithm performance (balanced accuracy) by dataset. Mean ± SD represents the average across all experimental configurations and 5 random seeds. Best indicates the peak performance achieved by the optimal configuration.

Dataset	LEM2		PRISM		RIPPER	
	Mean ± SD	Best	Mean ± SD	Best	Mean ± SD	Best
Breast Cancer	0.92 ± 0.02	0.94	0.89 ± 0.04	0.92	0.90 ± 0.03	0.94
Ckd	0.97 ± 0.04	0.99	0.61 ± 0.16	0.93	0.98 ± 0.03	1.00
Cleveland Heart	0.78 ± 0.04	0.82	0.54 ± 0.05	0.61	0.74 ± 0.05	0.78
Parkinsons	0.81 ± 0.08	0.86	0.68 ± 0.10	0.79	0.67 ± 0.09	0.78
Pima Diabetes	0.63 ± 0.05	0.69	0.59 ± 0.05	0.63	0.66 ± 0.03	0.70
Overall average	0.82 ± 0.13	-	0.66 ± 0.15	-	0.79 ± 0.14	-

the Chronic Kidney Disease dataset. On Breast Cancer dataset, federated LEM2 and RIPPER achieved equally high scores (0.94), outperforming PRISM (0.90). For the more challenging datasets such as Pima Indians Diabetes, federated RIPPER and LEM2 also demonstrated superior performance (0.70 and 0.69, respectively) compared to PRISM (0.63). The Cleveland Heart Disease and Parkinsons datasets showed similar trends, with federated LEM2 and RIPPER consistently outperforming federated PRISM.

5.2 Impact of Discretization Granularity

The number of discretization bins constitutes a crucial hyperparameter in learning rule-based models, since it determines the vocabulary of rule conditions which directly translates to the expressiveness versus complexity trade-off.

For federated LEM2, balanced accuracy shows a high level of stability between 3 and 6 bins (0.81–0.85) and decreasing at 2 (0.78-0.80). This suggests that 2 bins may not provide sufficient feature resolution for the algorithm to fully delineate class boundaries. Federated RIPPER achieved its most stable results with 4 bins (0.79-0.80). However, at higher bins granularities, its performance began to fluctuate and degrade, suggesting that an increasing number of bins, particularly with relatively small datasets, increases the risk of overfit-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

ting for this algorithm. This observation is generally consistent with its pruning strategy's tendency to lose effectiveness with significant expansion of feature and rule spaces. Federated PRISM achieved a peak at 2 and 4 bins (0.74) while displaying the most irregular pattern of results. Its rule building strategy may be sensitive to the concrete choice of discretization granularity leading to significant fluctuations across configurations (see Figure 1).

5.3 Scalability with Increasing Number of Clients

We also examined how performance scales in relation to the number of clients. Federated LEM2 achieved the highest stability with an increasing number of clients, reaching its maximum value with 5 clients and exhibiting low variance of results. This may indicate that the strategy of this model is able to effectively leverage diversified local data patterns even if each client has a reduced number of training data.

Federated RIPPER generally performed slightly worse than LEM2, characterized by higher variance of results depending on the number of clients, suggesting that its scalability depends on the discretization granularity. For example, with 4 bins, it remains stable across 2–5 clients, but with 5 bins, there is a substantial drop in performance (from 0.82 for 2 to 0.77 for 5 clients). This may be related to the fact that this algorithm needs an appropriately large amount of data to be able to effectively perform optimization and pruning of local rules. Too much fragmentation of data across many clients can lead to a decrease in the quality of these local rulesets negatively affecting the generalization ability of the aggregated global model.

Federated PRISM displayed the highest sensitivity to the increase in the number of clients which manifests itself as a general trend of decreasing balanced accuracy with increasing clients count regardless of bins granularity. This seems consistent with its separate-and-conquer rule building strategy, which requires an appropriate number of examples per class. When the number of clients increases and the number of available local training data decreases, federated PRISM does not handle the extraction of reliable rules well, in particular for the minority class in imbalanced datasets.

5.4 Benefits of Global Aggregation

A fundamental question in the case of rule-based FL is whether the federated model is able to deliver performance approaching the centralized one's while also providing genuine advantages over separate local models trained in isolation by each client. Table 2 presents a comparison of global model performance against both averaged balanced accuracy of local and centralised models on the global test set. It shows that global models consistently outperformed the average local models. Notably, federated LEM2 exceeds centralized training (red) while federated PRISM and RIPPER achieve statistical parity.

All three algorithms exhibited consistent and highly significant increase in prediction quality due to global aggregation of rules in comparison with averaged

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

Table 2. Comparison of learning paradigms: local models, federated global models and centralized baselines. Statistical significance assessed via paired Wilcoxon signed-rank tests across 5 random data splits per dataset (n=25 paired observations). Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Algorithm	Avg Local	Fed. Global	Cent.	Global vs Local	Global vs Cent.
LEM2	0.752	0.821	0.798	+9.1% ($p < 0.001$)***	+2.8% ($p = 0.011$)*
PRISM	0.614	0.662	0.685	+7.8% ($p = 0.004$)**	-3.4% ($p = 0.692$)
RIPPER	0.740	0.792	0.810	+7.0% ($p < 0.001$)***	-2.2% ($p = 0.474$)

local models. Federated LEM2 achieved the largest increase (9.1%), improving from 0.752 to 0.821. Federated PRISM exhibited a performance increase equal to 7.8%, achieving an improvement from 0.614 for average local models to 0.662 for the global model. Also, federated RIPPER benefited from global aggregation, achieving 7.0% relative improvement in prediction quality.

Compared to its centralized baseline, federated LEM2 significantly outperformed it (0.821 versus 0.798). In turn, both federated PRISM and RIPPER achieved statistical parity with centralized models. Both federated PRISM and RIPPER did not differ significantly from the centralized baseline ($p = 0.692$ and $p = 0.474$, respectively).

5.5 Rulesets Characteristics and Interpretability

Table 3 contains a summary of rulesets extracted by the three algorithms across all experimental configurations and datasets averaged over 5 random data splits.

6 Discussion

The aim of this work was to verify whether rule-based FL can be effectively applied to small biomedical datasets. The presented results provide an affirmative answer (average balanced accuracy from 0.66 to 0.82 across algorithms and datasets). The coverage analysis shows that the models provide alternative interpretable results with either high coverage of cases or compact and very precise set of rules, maintaining comparable performance in both cases. Global aggregation moreover allows to consistently outperform averaged results of local models achieving 7.0-9.1 % performance gain. Remarkably, federated LEM2 performed better than the centralized baseline ($p = 0.011$), while federated PRISM and RIPPER achieved statistical parity with centralized versions, suggesting that privacy-enchancing federated rule-based learning does not necessarily entail a significant loss in prediction quality.

Federated LEM2 and RIPPER achieved similarly high performance (on average 0.82 and 0.79 respectively). The picture that emerges from the analysis of federated PRISM results is more complicated. It achieved significantly lower average performance on the Cleveland Heart Disease data which can be explained by the markedly low coverage (31%), which means that as many as 69% of cases

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

Table 3. Ruleset characteristics by dataset and algorithm. Per-dataset cells are mean ± SD over all bin and client settings and five random seeds. The All datasets row shows the overall mean ± SD for a given algorithm across all experiments.

Dataset	Algorithm	Total rules	Conds. per rule	Coverage (%)	Jaccard sim.
Breast Cancer	LEM2	43.2 ± 6.8	2.42 ± 0.29	96.63 ± 10.31	0.02 ± 0.02
	PRISM	15.0 ± 2.4	4.66 ± 4.13	93.76 ± 4.19	0.07 ± 0.06
	RIPPER	12.8 ± 3.7	1.31 ± 0.18	42.93 ± 6.11	0.11 ± 0.12
Ckd	LEM2	6.9 ± 1.4	1.08 ± 0.10	99.36 ± 1.52	0.21 ± 0.15
	PRISM	2.8 ± 1.3	1.39 ± 1.14	82.73 ± 9.22	0.40 ± 0.36
	RIPPER	4.9 ± 1.4	1.00 ± 0.00	26.53 ± 2.52	0.11 ± 0.17
Cleveland Heart	LEM2	51.5 ± 4.6	2.73 ± 0.40	99.21 ± 1.47	0.01 ± 0.01
	PRISM	9.8 ± 2.1	6.81 ± 1.16	31.21 ± 16.98	0.03 ± 0.05
	RIPPER	7.8 ± 1.6	1.68 ± 0.25	62.06 ± 10.66	0.04 ± 0.06
Parkinsons	LEM2	30.7 ± 5.4	1.88 ± 0.33	98.25 ± 3.82	0.04 ± 0.03
	PRISM	7.1 ± 1.3	4.96 ± 2.59	86.19 ± 8.23	0.03 ± 0.05
	RIPPER	6.5 ± 1.9	1.10 ± 0.13	80.37 ± 12.91	0.05 ± 0.07
Pima Diabetes	LEM2	133.4 ± 56.1	2.87 ± 0.28	84.89 ± 22.31	0.02 ± 0.02
	PRISM	13.8 ± 3.9	3.46 ± 1.31	58.33 ± 16.88	0.09 ± 0.10
	RIPPER	9.6 ± 3.5	1.87 ± 0.45	41.10 ± 14.14	0.10 ± 0.11
All datasets	LEM2	53.1 ± 49.8	2.20 ± 0.72	95.67 ± 12.40	0.06 ± 0.10
	PRISM	9.7 ± 5.1	4.26 ± 2.96	70.44 ± 25.97	0.12 ± 0.22
	RIPPER	8.3 ± 3.8	1.39 ± 0.42	50.60 ± 21.30	0.08 ± 0.12

were predicted using the default rule. It also achieved significantly worse results Chronic Kidney Disease and Parkinsons datasets despite high coverage. However, it achieved competitive results for Breast Cancer Wisconsin (with high coverage of 94 %) and to some extend on Pima Indians Diabetes. This leads to a conclusion that it is best suited for carefully selected datasets. LEM2's maximal generalization principle results in numerous rulesets with simple rules allowing for broad coverage of cases (96% on average), while RIPPER's aggressive pruning strategy achieves competitive results despite lower overall coverage (51%).

Regarding the optimal number of discretization bins, it is dependent on the dataset, but generally it is 4-5 bins in the case off datasets type considered in this work. Fewer bins (in particular 2) lead to insufficient feature resolution and limited ability to detect complex patterns. Too many bins increase the risk of overfitting and at the same time can lead to increased communication overhead. The greatest irregularity was presented by federated PRISM, and federated RIPPER showed significant degradation at higher granularity.

Both federated LEM2 and federated RIPPER exhibited satisfactory robustness to the increase in the number of clients. The results also illustrate that different algorithms set different thresholds of minimum amount of local data for effective training and federated PRISM seems to suffer the most from decreasing this amount. In general, these algorithms stand in opposition to neural networks, where usually a significantly larger number of clients is needed for effective training in federated settings [23].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

All three algorithms exhibited consistent increase in prediction quality beyond a mere ensemble averaging, indicating that the detected local patterns are indeed highly complementary. Federated LEM2 exhibited the largest gain (9.1%) demonstrating the complementarity of the maximal generalization principle across clients. Similar results were also obtained by federated PRISM (7.8%) and federated RIPPER (7.0%). The obtained average Jaccard similarity index values (0.06-0.12) indicating that only 6-12% of rules overlap between clients confirm this. These results align with the PAC learning theory and could be related to the fact that aggregation creates a more expressive hypothesis space that combines local approximations of the underlying target function offering richer representation than any of the local or even centralized homogeneous models [32]. Additionally, the fact that some pairs of clients share rules while others do not, as indicated by the standard deviation values of the similarity metric, indicate data heterogeneity across clients.

All considered algorithms provide rulesets with different profiles that can all be inspected by a human. Federated RIPPER provides the most compact, practical, memorizable and actionable rules offering an effective default class. It handles many cases using the default majority rule, but offers high precision for matched non-default rules. This shows that the algorithm balances well a small number of well-chosen explicit rules with the default rule, achieving high prediction quality despite moderate coverage.

Federated LEM2 provides a more complete picture by prioritizing greater coverage with larger number of rules, which however are simple. The high average coverage value suggests that this algorithm is almost always able to provide a precise explanation of its predictive decision tracking it back to a concrete non-default rule.

Federated PRISM merges compactness with complexity offering a detailed insight into predictions, provided that the coverage is sufficient. It offers better average coverage than federated RIPPER, while demonstrating the highest variability of coverage. Hovewer, its rules are on average less accurate, and the coverage value is strongly correlated with predictive performance.

From this emerges the conclusion that different algorithms can be appropriate for different clinical contexts (e.g., federated RIPPER for quick decision support with reliable defaults, federated LEM2 for comprehensive diagnosis covering almost all case with explicit rules). The results indicate that given moderately complex datasets allowing to achieve a sufficiently high coverage, federated PRISM also achieves satisfactory prediction quality, providing the most nuanced (in terms of the number of conditions) explanations of its predictions.

Building on these findings, real-world deployment requires careful considerations. Manual validation of rules is essential to detect spurious patterns. Furthermore, by implementing Outcome-Action Pairing (OAP) [2], predictions should be translated to actionable clinical recommendations by leveraging the interpretability of rules to increase clinician trust and enable the effective use of clinical decision support.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

7 Conclusions and Future Work

Our results demonstrate that federated rule learning can be a viable approach for handling sensitive medical tabular data. The federated versions of the models achieved results competitive with those obtained by the centralized versions and outperformed isolated local ones, highlighting the benefits of collaborative learning. In addition, the federated LEM2 achieved a consistently higher average balanced accuracy across the tested configurations than the model trained in a centralised way.

The interpretability of such an approach is particularly suited for high-stakes clinical domains that require regulatory approval. Effective operation with a small number of clients and a small amount of data also makes it an interesting option for rare disease research, where data can be scarce. Rule-based FL is thus a complementary approach to black-box FL, demonstrating that rule-based FL can provide both transparency and competitive performance for real-world clinical applications.

This work has several limitations: (i) only small biomedical datasets were used; (ii) solely union aggregation was explored; (iii) only equal-width discretization was considered; (iv) experiments were limited to at most 5 clients; and (v) random data partitioning with an equal load for each client was used, whereas real-world institutional applications may contain varying numbers of data samples and introduce batch effects [20].

Regarding privacy, clients only send the minimum and maximum values for each feature to the server. These values cannot be linked to specific patients without additional external knowledge. They also carry no information about the other features of patients, meaning no individual patient records are shared. While this provides meaningful protection against re-identification, it falls short of the guarantees offered by formal differential privacy, which would also obscure whether any particular value exists in the dataset. Adding calibrated noise to these values before transmission is a possible step toward formal differential privacy, which we identify as a direction for future work.

Future work should address these limitations by analyzing other aggregation and discretization methods (e.g., abstractions [17]), assessing scalability on larger datasets with more clients, comparing against federated boosting and neural networks and evaluating the approach in realistic clinical environments with medical expert assessment of resulting rulesets.

Acknowledgments. This publication is partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement "Sano" No. 857533, and by the "Sano" pro-ject, carried out within the framework of the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13, co-financed by the European Regional Development Fund. The publication was created within the project of the Minis-ter of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

Disclosure of Interests. The authors declare no competing interests.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

References

1. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., de Gusmão, P.P.B., Lane, N.D.: Flower: A friendly federated learning research framework (2022), https://arxiv.org/abs/2007.14390
2. Bolivar, L.E.M., Satzer, M.: Essential concepts in artificial intelligence: A guide for pediatric providers. Children **12**, 1386 (10 2025). https://doi.org/10.3390/CHILDREN12101386, https://pmc.ncbi.nlm.nih.gov/articles/PMC12563249/
3. Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science **16** (2001), https://api.semanticscholar.org/CorpusID:62729017
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. 2010 20th International Conference on Pattern Recognition pp. 3121–3124 (2010), https://api.semanticscholar.org/CorpusID:11557689
5. Cendrowska, J.: Prism: An algorithm for inducing modular rules. International Journal of Man-Machine Studies **27**(4), 349–370 (1987)
6. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning. pp. 115–123 (1995)
7. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (Dec 2006)
8. Fu, Y., Wang, D., Xiong, H., Liu, K.: Tabular data-centric ai: Challenges, techniques and future perspectives. International Conference on Information and Knowledge Management, Proceedings pp. 5522–5525 (10 2024). https://doi.org/10.1145/3627673.3679102, https://dl.acm.org/doi/10.1145/3627673.3679102
9. Fürnkranz, J., Gamberger, D., Lavrač, N.: Foundations of rule learning. In: Cognitive Technologies (2012), https://api.semanticscholar.org/CorpusID:12596492
10. Fürnkranz, J.: Separate-and-conquer rule learning. Artificial Intelligence Review **13**(1), 3–54 (1999)
11. Fürnkranz, J., Widmer, G.: Incremental reduced error pruning. In: International Conference on Machine Learning (1994), https://api.semanticscholar.org/CorpusID:5310845
12. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision making and a "right to explanation". AI Magazine **38**(3), 50–57 (Sep 2017). https://doi.org/10.1609/aimag.v38i3.2741, http://dx.doi.org/10.1609/aimag.v38i3.2741
13. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? In: Advances in Neural Information Processing Systems. vol. 35, pp. 507–520 (2022)
14. Grzymala-Busse, J.W.: Lers–a system for learning from examples based on rough sets. In: Intelligent Decision Support, pp. 3–18. Springer (1992)
15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys **51**(5), 1–42 (2018)
16. Heusinger, M., Raab, C., Rossi, F., Schleif, F.M.: Federated learning - methods, applications and beyond. In: ESANN 2021 proceedings. p. 1–10. ESANN 2021, Ciaco - i6doc.com (2021). https://doi.org/10.14428/esann/2021.es2021-4, http://dx.doi.org/10.14428/esann/2021.ES2021-4

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38

17. Ibias, A., Capala, K., Varma, V.R., Drozdz, A., Sousa, J.: Improving noise robustness through abstractions and its impact on machine learning (2024)
18. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart Disease. UCI Machine Learning Repository (1989), DOI: https://doi.org/10.24432/C52P4X
19. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., et al.: Advances and open problems in federated learning. Foundations and Trends in Machine Learning **14**(1-2), 1–210 (2021)
20. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)
21. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems. vol. 2, pp. 429–450 (2020)
22. Little, M.: Parkinsons. UCI Machine Learning Repository (2007), DOI: https://doi.org/10.24432/C59C74
23. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
24. Michalski, R.S.: A theory and methodology of inductive learning. Artif. Intell. **20**, 111–161 (1983), https://api.semanticscholar.org/CorpusID:58987397
25. Pawlak, Z.: Rough sets. International Journal of Computer & Information Sciences **11**(5), 341–356 (1982)
26. Pilgram, L., Ko, H., Tung, A., Emam, K.E.: Protecting patient privacy in tabular synthetic health data: a regulatory perspective. NPJ digital medicine **8** (12 2025). https://doi.org/10.1038/S41746-025-02112-0, https://pubmed.ncbi.nlm.nih.gov/41315669/
27. Rubini, L., Soundarapandian, P., , Eswaran, P.: Chronic Kidney Disease. UCI Machine Learning Repository (2015), DOI: https://doi.org/10.24432/C5G020
28. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
29. Smith, J., Everhart, J., Dickson, W., Knowler, W., Johannes, R.: Pima Indians Diabetes Database. UCI Machine Learning Repository (1988), https://search.r-project.org/CRAN/refmans/mlbench/html/PimaIndiansDiabetes.html
30. Tan, P.N., Steinbach, M., Karpatne, A., Kumar, V.: Introduction to Data Mining (2nd Edition). Pearson, 2nd edn. (2018)
31. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: Contextualizing explainable machine learning for clinical end use (2019), https://arxiv.org/abs/1905.05134
32. Valiant, L.G.: A Theory of the Learnable. Communications of the ACM **27**(11), 1134–1142 (1984). https://doi.org/10.1145/1968.1972
33. Wang, D., Huang, Y., Ying, W., Bai, H., Gong, N., Wang, X., Dong, S., Zhe, T., Liu, K., Xiao, M., Wang, P., Wang, P., Xiong, H., Fu, Y.: Towards data-centric ai: A comprehensive survey of traditional, reinforcement, and generative approaches for tabular data transformation (2025), https://arxiv.org/abs/2501.10555
34. Wolberg, W., Mangasarian, O., Street, N., Street, W.: Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository (1993), DOI: https://doi.org/10.24432/C5DW2B
35. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications (2019), https://arxiv.org/abs/1902.04885
36. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_38