FLARE: RELIABLE LOCAL EXPLANATIONS THROUGH FEDERATED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative model training without sharing raw data, but local explanation methods remain constrained by the data available at individual clients. When local data are sparse, imbalanced, or heterogeneous, query neighborhoods may provide insufficient support for representative and faithful explanations. Although federated explainability methods can exploit distributed information in different ways, the use of collaboratively learned representations to improve real-instance, query-specific neighborhood construction for local surrogate explanations remains largely unexplored. We propose **FLARE**, a federated framework for reliable local explanations. To the best of our knowledge, FLARE is the first framework to combine federated representation learning with client-local, real-instance neighborhoods for query-specific surrogate explanations. A collaboratively trained SCARF encoder guides local source–target neighborhood construction, while support-aware neighborhood completion and adaptive surrogate weighting improve reliability. Explanations are generated using weighted Ridge surrogates and evaluated for fidelity, support, stability, and robustness. We evaluate FLARE across five tabular datasets spanning seven classification tasks, including predefined source–target domain shifts and natural multi-client partitions, under varying client counts, federated optimizers, and black-box predictors. FLARE achieves the strongest overall fidelity across the evaluated baselines, with sensitivity analyses showing that it remains effective across different client counts and black-box predictors. These results show that cross-client structural knowledge can improve local explanation construction while keeping raw data and explanation generation client-local.

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) is becoming increasingly important as complex machine learning models are deployed in high-stakes domains such as healthcare, finance, and public decision support Mersha et al. (2024); Sahoh & Choksuriwong (2023). For example, a pneumonia-risk model learned that asthma was associated with lower mortality because asthmatic patients had received more intensive care. Such a misleading association could lead the model to underestimate the risk of future patients Caruana et al. (2015). This example illustrates why explanations are important for identifying potentially harmful associations before machine learning models are deployed in high-stakes settings.

Recent work on local explanations includes LIME, which explains an individual prediction by fitting an interpretable surrogate in its local neighborhood Ribeiro et al. (2016); Bharati et al. (2023), as well as methods that improve how this neighborhood is constructed. Conventional random perturbations may violate feature dependencies, move away from the underlying data distribution, and weaken the resulting explanation Laugel et al. (2018); Tan et al. (2023). ITL-LIME addresses this limitation in low-resource settings by combining real source-domain instances with nearby target-domain observations and using learned representations to weight their relevance Raza et al. (2025). However, this approach assumes that the data required for representation learning and neighborhood construction can be accessed jointly. In federated settings, where data remain distributed across separate clients, local explanations therefore need a way to benefit from broader data structure without pooling raw records.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Federated learning enables collaboration across distributed clients by exchanging model updates rather than raw records Blanco-Justicia et al. (2021). Existing federated explainability approaches commonly apply post-hoc methods such as LIME to models trained through federated learning Patel et al. (2024). In such approaches, however, federated collaboration primarily supports predictive model training, while the neighborhood used for a local surrogate explanation is still constructed independently at each client. This limitation becomes particularly important when a client has limited, sparse, or imbalanced observations around a query, making it difficult to construct a representative local neighborhood. Although recent federated explainability methods have begun to collaboratively construct auxiliary information for other explainers, such as representative background data for SHAP Ducange et al. (2026a), the use of federated representation learning to guide real-instance neighborhood construction for local surrogate explanations remains largely unexplored. Such a mechanism can allow each client to benefit from similarity structure learned across the federation when identifying relevant local instances, improving neighborhood quality under heterogeneous or data-limited conditions without exchanging raw records. To the best of our knowledge, FLARE is the first framework to combine federated representation learning with client-local, real-instance neighborhood construction for query-specific surrogate explanations.

Specifically, FLARE learns a shared representation across clients and uses it to guide query-specific neighborhood construction from real local instances. This allows cross-client structural information to support local explanations without exchanging raw data, while black-box interaction and surrogate fitting remain client-local. FLARE further adapts neighborhood composition and surrogate weighting to improve support around the query and its local decision boundary. The main contributions of this work are: (i) To the best of our knowledge, **FLARE** is the first framework to combine federated representation learning with client-local, real-instance neighborhood construction for query-specific surrogate explanations, while keeping raw data, black-box interaction, surrogate fitting, and explanation generation local. (ii) A federated SCARF representation-learning mechanism, optimized with FedAvg or FedProx, that provides a shared latent space for selecting and weighting relevant source–target instances. (iii) Two reliability mechanisms: *support-aware neighborhood completion*, which augments under-supported neighborhoods with nearby real counter-class evidence, and *adaptive surrogate weighting*, which adjusts instance influence according to locality and class support. (iv) A reliability-oriented evaluation protocol combining conventional and held-out fidelity with neighborhood support, decision-boundary performance, stability, and local robustness. (v) An extensive evaluation across five tabular datasets spanning seven classification tasks, under multiple client configurations, federated optimizers, and black-box predictors.

To support reproducibility, the code, data, and other resources used in this study are made available through anonymous repository[1].

2 Related Work

Our related work encompasses two main areas: local post-hoc explanation methods and explainability in federated learning.

Local Post-hoc Explainability Local post-hoc methods such as LIME Ribeiro et al. (2016) explain individual predictions by fitting an interpretable surrogate around a target instance. Their reliability depends strongly on neighborhood construction and weighting. Random perturbations may violate feature dependencies, leave the data manifold, and provide poor coverage of the local decision region Laugel et al. (2018). Subsequent methods therefore improve sampling and locality. GLIME Tan et al. (2023), for example, revisits the sampling distribution to improve locality and stability. More recently, instance-based approaches use real observations instead of synthetic perturbations. ITL-LIME Raza et al. (2025) combines neighboring target observations with relevant source-domain instances and uses a learned representation to determine locality.

Explainability in Federated Learning Most explainable federated learning approaches apply established post-hoc methods such as LIME or SHAP to federatively trained models Biswas et al. (2025); Kalakoti et al. (2025); Tunduny & Shibwabo (2026); Rani (2025). Other methods explicitly collaborate over explanation-related information. Abtahi et al. Abtahi Fahliani et al. (2024) aggre-

[1] https://anonymous.4open.science/r/FLARE-B858/README.md

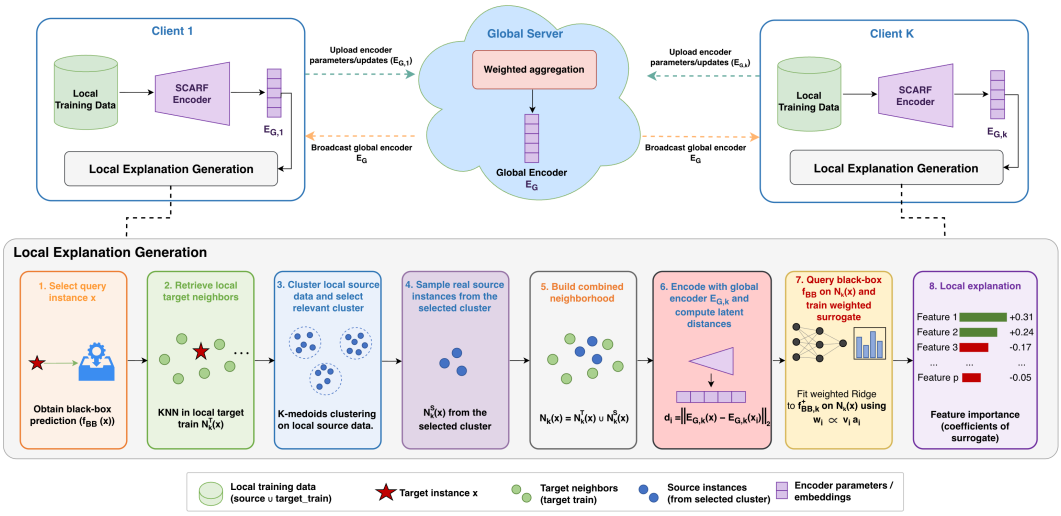


Figure 1: **Overview of FLARE.** Each client trains a SCARF encoder using FedAvg or FedProx producing a shared encoder E_G, which is broadcast to all clients; with client-local Batch Normalization, client k uses $E_{G,k}$. After T rounds, the encoder is frozen for explanation generation. For a query x, the client: (1) obtains $f_{\mathrm{BB}}(x)$; (2) retrieves target neighbors $N_k^T(x)$; (3) clusters the local source data and selects a relevant cluster; (4) samples source instances $N_k^S(x)$; (5) constructs $\mathcal{N}_k(x) = N_k^T(x) \cup N_k^S(x)$; (6) computes latent distances $d_i = \|E_{G,k}(x) - E_{G,k}(x_i)\|_2$; (7) queries f_{BB} on $\mathcal{N}_k(x)$ and fits a weighted Ridge surrogate using $w_i \propto v_i a_i$; and (8) returns the surrogate coefficients as the local explanation.

gate locally generated SHAP- and DeepLIFT-based interpretability information, while Federated SHAP Ducange et al. (2026b) constructs representative background data using federated clustering or generative modeling. Haffar et al. Haffar et al. (2023) (FedRF-Surrogate) restricted-depth random forests as interpretable surrogates of federated black-box models, while Bonsignori et al. Bonsignori et al. (2026) propose FastSHAP++, which aggregates locally computed SHAP explanations across clients. Despite these advances, federated collaboration is primarily used to train predictive models, aggregate explanations, or construct auxiliary reference information; the local neighborhood on which a query-specific surrogate explanation depends remains largely client-local and is not informed by collaboratively learned similarity structure. This limitation is particularly important when local data are sparse, imbalanced, or heterogeneous, where the available observations may provide insufficient evidence around a query and its decision boundary. FLARE addresses this gap by using federated representation learning specifically to guide real-instance neighborhood construction, thereby transferring cross-client structural knowledge into local explanation generation without exchanging raw observations. Support-aware neighborhood completion and adaptive surrogate weighting further improve reliability when local evidence is insufficient.

3 Methodology

We propose **FLARE**, a federated framework for generating reliable local explanations in distributed and heterogeneous settings. Figure 1 illustrates the conceptual framework, while Algorithm 1 summarizes the complete workflow. FLARE is organized into two stages: collaborative representation learning and query-specific explanation generation. In the first stage, clients collaboratively train a SCARF representation model while keeping their raw observations local. In the second stage, the resulting federated encoder is frozen and used to define representation-space locality for constructing and assessing query-specific explanations. The remainder of this section describes the individual steps of this two-stage architecture in detail.

Algorithm 1 FLARE

Require: $\{D_k^S, D_k^{T,\text{train}}, D_k^{T,\text{test}}, f_{\text{BB}}\}_{k=1}^K$, where f_{BB} is trained on $D_k^{T,\text{train}}$, rounds T
Ensure: Client-local explanations
Federated representation learning
Initialize SCARF model $F_{\omega^{(0)}} = h_{\phi^{(0)}} \circ E_{\theta^{(0)}}$
for $t = 0, \ldots, T-1$ **do**
  Select $\mathcal{S}_t$ and broadcast shared parameters
  for all $C_k \in \mathcal{S}_t$ **in parallel do**
    Train SCARF on $D_k^S \cup D_k^{T,\text{train}}$, adding the FedProx penalty when enabled
    Return shared parameters
  end for
  Aggregate parameters by sample-size-weighted averaging
end for
Construct and freeze $E_{G,k}$, retaining local BatchNorm
Client-local explanation generation
for all (C_k, x) with $x \in D_k^{T,\text{test}}$ **do**
  Retrieve target neighbors $N_k^T(x)$
  Select source instances $N_k^S(x)$ using K-medoids, proximity, and a class-mismatch penalty
  Form $\mathcal{N}_k(x) = N_k^T(x) \cup N_k^S(x)$
  Compute latent-space locality weights using $E_{G,k}$
  Query f_{BB} on $\mathcal{N}_k(x)$ and assess class support
  if support-aware completion is enabled and support is insufficient **then**
    Add admissible counter-class evidence through constrained replacement
  end if
  Compute adaptive fitting weights when enabled
  Fit weighted Ridge to the black-box probabilities
  Return its coefficients and compute reliability metrics
end for

3.1 Federated Setting and Representation Learning

We consider K clients $\mathcal{C} = \{C_1, \ldots, C_K\}$. Each client holds an auxiliary *source domain* and a *target domain* associated with the predictive task. The source domain may contain historical or reference observations available at the client, whereas the target domain represents the deployment population for which predictions and explanations are required. Formally, client C_k holds private source data D_k^S and target data D_k^T where $D_k^T = D_k^{T,\text{train}} \cup D_k^{T,\text{test}}$. The client-local black-box predictor $f_{\text{BB}} : \mathcal{X} \to \mathcal{Y}$ is trained exclusively on $D_k^{T,\text{train}}$. The held-out set $D_k^{T,\text{test}}$ is excluded from black-box and representation training and is used only for predictive evaluation and as a source of explanation queries.

Before federated training begins, the server initializes a SCARF model $F_\omega = h_\phi \circ E_\theta$, where $\omega = (\theta, \phi)$ comprises the parameters of the encoder and projection head. Federated representation learning then proceeds for T communication rounds using the clients' local encoder-training pools. For each observation $x \in D_k^{\text{enc}}$, SCARF constructs a corrupted view $\tilde{x}$ by replacing selected numerical values and categorical blocks with values sampled from other observations in the same local minibatch. The encoder E_θ maps the original and corrupted observations to latent representations, and the projection head h_ϕ maps these representations into the contrastive space. The SCARF objective encourages agreement between the original and corrupted views of the same observation while separating representations of different observations.

Let $\omega = (\theta, \phi)$ denote the trainable parameters of the SCARF encoder and projection head. At communication round t, the server broadcasts the current shared parameters $\omega^{(t)}$ to the participating clients. Each selected client optimizes the SCARF objective over D_k^{enc} and returns its updated model parameters. Under FedAvg, the server computes

$$\omega^{(t+1)} = \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} \omega_k^{(t+1)}, \qquad n_k = |D_k^{\text{enc}}|, \tag{1}$$

where $\mathcal{S}_t$ denotes the set of clients participating in round t. To reduce client drift under heterogeneous local distributions, FLARE also supports FedProx. Its local objective is

$$\mathcal{L}_k^{\text{prox}} = \mathcal{L}_k^{\text{SCARF}} + \frac{\mu}{2} \left\| \omega - \omega^{(t)} \right\|_2^2. \tag{2}$$

FLARE retains each client's Batch Normalization parameters and running statistics locally, while the remaining SCARF parameters are shared and aggregated. Only these shared parameters are included in the FedProx proximal term. After T communication rounds, client C_k combines the

globally aggregated parameters with its local Batch Normalization state to construct the frozen explanation encoder $E_{G,k}$. The projection head is used only during contrastive training and is discarded before explanation generation.

3.2 Client-Local Source–Target Neighborhood Construction

Source observations provide complementary real-instance evidence and are combined with target-training neighbors to construct each query-specific neighborhood. For a query $x \in D_k^{T,\text{test}}$, FLARE constructs the neighborhood using only real observations available at client C_k. Target-side evidence consists of the K_T Euclidean nearest neighbors of x in the preprocessed target-training data, denoted by $N_k^T(x) \subset D_k^{T,\text{train}}$. Consequently, other target-test observations cannot enter the explanation neighborhood. For source-side evidence, FLARE partitions D_k^S into M K-medoids clusters $\{G_{k,1}, \ldots, G_{k,M}\}$ in the preprocessed feature space. Let $m_{k,j}$ denote the medoid of cluster $G_{k,j}$. The local black box predicts the class of query instance:

$$\hat{c}_x = \mathbb{I}\big[f_{\text{BB}}^+(x) \geq 0.5\big]\,, \tag{3}$$

where $f_{\text{BB},k}^+(x)$ denotes the local black-box model's predicted probability that x belongs to the positive class. For each source cluster $G_{k,j}$, let $m_{k,j}$ denote its medoid and $\hat{c}_{k,j}$ the majority black-box class of its members. FLARE assigns each cluster the query-dependent selection score

$$s_{k,j}(x) = \|x - m_{k,j}\|_2 + \eta\, \mathbb{I}(\hat{c}_{k,j} \neq \hat{c}_x)\,. \tag{4}$$

Here $\|x - m_{k,j}\|_2$ measures the distance between the query and the cluster medoid and $\eta\, \mathbb{I}(\hat{c}_{k,j} \neq \hat{c}_x)$ adds a class-mismatch penalty η when the cluster's majority class differs from the query's predicted class. We set $\eta = 10^6$, thereby prioritizing the nearest class-compatible cluster and falling back to the nearest source cluster when no compatible cluster exists. FLARE samples at most n_S observations without replacement from the selected cluster to construct $N_k^S(x)$. The initial local explanation neighborhood is $\mathcal{N}_k^{(0)}(x) = N_k^T(x) \cup N_k^S(x)$. The query x guides target-neighbor retrieval, source-cluster selection, and locality weighting, but is not included as a surrogate-training observation in $\mathcal{N}_k^{(0)}(x)$.

3.3 Reliability-Aware Local Surrogate

The frozen federated encoder measures the relevance of every neighborhood observation $x_i \in \mathcal{N}_k(x)$ to the query. FLARE computes

$$d_i = \|E_{G,k}(x) - E_{G,k}(x_i)\|_2\,, \qquad v_i = \exp\left(-\frac{d_i^2}{2\sigma^2}\right), \tag{5}$$

where v_i is the latent-locality weight and σ controls the kernel width. Observations closer to the query in the federated representation space therefore receive greater influence. The client-local black box supplies the positive-class probability $p_i = f_{\text{BB},k}^+(x_i)$ and the corresponding hard prediction $y_i = \mathbb{I}[p_i \geq 0.5]$.

FLARE assesses whether the initial neighborhood contains sufficient evidence from both black-box classes. For each class $c \in \{0, 1\}$, FLARE computes its raw and locality-weighted support:

$$r_c = \frac{\sum_i \mathbb{I}[y_i = c]}{|\mathcal{N}_k^{(0)}(x)|}, \qquad s_c = \frac{\sum_i v_i \mathbb{I}[y_i = c]}{\sum_i v_i}. \tag{6}$$

Support-aware completion is triggered when the raw or locality-weighted support of the weaker class falls below threshold τ_s.

Completion candidates are drawn exclusively from $D_k^S \cup D_k^{T,\text{train}}$ and are assigned black-box predictions rather than ground-truth labels. FLARE considers the nearest real candidates predicted as the under-supported class and replaces the farthest observations belonging to the locally dominant class. The neighborhood size therefore remains unchanged. An admissible candidate must lie within a radius defined by the q-quantile of the original latent distances, expanded by factor γ. It must also be no farther than γ times the distance of the majority-class observation it replaces. At most a fraction ρ of the neighborhood can be replaced. We use $\tau_s = 0.10, q = 0.95, \gamma = 1.25, and \rho = 0.15$ across all datasets.

FLARE subsequently applies adaptive class-support weighting when both black-box classes contain at least the required minimum number of real neighborhood observations. Let $M_c = \sum_{i:y_i=c} v_i$ denote the locality mass of class c, and let $c^- = \arg\min_c M_c$ be the class with lower locality mass. Its multiplier is

$$a = \min\left(\frac{M_{1-c^-}}{\max(M_{c^-}, \epsilon)}, a_{\max}\right), \tag{7}$$

where $a_{\max}$ limits excessive reweighting. The adaptive multiplier and final fitting weights are

$$a_i = \begin{cases} a, & y_i = c^-, \\ 1, & \text{otherwise}, \end{cases} \qquad w_i = a_i v_i \frac{\sum_j v_j}{\sum_j a_j v_j}. \tag{8}$$

If either class has insufficient support, adaptive reweighting is not applied and $w_i = v_i$. In our implementation, the minimum class count is five and $a_{\max} = 5$. Using these fitting weights, FLARE learns a weighted Ridge surrogate in the feature space:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_i w_i \left(p_i - \beta_0 - \boldsymbol{\beta}^\top x_i\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2. \tag{9}$$

The coefficient vector $\boldsymbol{\beta}$ constitutes the query-specific explanation. Adaptive class-support weights affect only surrogate fitting; all reported fidelity metrics are computed using the original locality weights v_i. Thus, adaptive weighting cannot directly increase the reported fidelity by modifying its evaluation weights.

3.4 Explanation Reliability Assessment

FLARE evaluates fidelity, generalization, support, stability, and robustness. Locality-weighted F1, AUC, MCC, and balanced accuracy measure surrogate agreement with the black box within the neighborhood, unlike predictive metrics that compare black-box outputs with target-test labels. Held-out fidelity assesses local generalization through repeated stratified neighborhood splits: a temporary surrogate is fitted on each training fold and evaluated against black-box outputs on the unseen fold, while the final explanation uses the complete neighborhood.

The *support rate* is the fraction of neighborhoods containing both black-box classes, meeting the minimum class count, and satisfying the raw and locality-weighted minority-support thresholds. An explanation is *reliable* if it is supported, has a non-degenerate surrogate and valid held-out evaluation, and achieves held-out balanced accuracy ≥ 0.70 and MCC ≥ 0.50. When evaluated, JC must also be at least 0.50; LLE does not affect reliability.

Boundary fidelity is measured using MCC for queries satisfying $0.40 \leq f^+_{\mathrm{BB},k}(x) \leq 0.60$. Jaccard consistency (JC) measures the mean pairwise overlap between top-k feature sets across repeated explanations. The local Lipschitz estimate (LLE) measures sensitivity to valid local perturbations Alvarez-Melis & Jaakkola (2018):

$$\mathrm{LLE}(x) = \max_{x' \in \mathcal{P}_k(x)} \frac{\|\mathcal{E}_k(x) - \mathcal{E}_k(x')\|_2}{d_{\mathcal{X}}(x, x')}, \tag{10}$$

where $\mathcal{P}_k(x)$ is the perturbation set, $\mathcal{E}_k(\cdot)$ is the normalized semantic-feature importance vector, and $d_{\mathcal{X}}$ is the mixed-type input distance. Perturbations use small numerical changes and valid categorical substitutions from nearby target-training data, with the explanation procedure rerun for each perturbation. Higher JC and lower LLE indicate better stability and robustness, respectively.

4 Experimental Setup

4.1 Benchmark Datasets

We evaluate FLARE on seven binary-classification datasets: Adult, Bank, Student–STD1 , Student–STD2, Diabetes–DTD1, Diabetes–DTD2, and ACS Income. Their characteristics and domain constructions are reported in Appendix Table 6. Adult and Bank use a 75/25 source–target split. Student Depression uses Western India as the source and Eastern India (STD1) or Southern India (STD2) as the target, while Diabetes uses mainland U.S. observations as the source and the U.S. Virgin Islands (DTD1) or Guam (DTD2) as the target. For these six datasets, observations are

randomly distributed among 3, 5, and 10 clients using balanced, label-stratified partitions. For naturally non-IID evaluation, ACS Income assigns one client to each of the 50 U.S. states and applies a 75/25 source–target split within each state. For all datasets, each target partition is further divided into 80% training and 20% test sets.

4.2 Benchmark Baselines

We compare FLARE with three federated explainability baselines using identical data partitions, preprocessing, and target-test queries.

Fed-LIME combines a FedAvg-trained black box with client-local LIME surrogates fitted to perturbed samples Rani (2025).

FedRF fits client-local random-forest surrogates to federated black-box predictions and derives explanations from prediction-consistent trees Haffar et al. (2023).

FastSHAP++ federatively trains a masking surrogate and an amortized SHAP explainer Bonsignori et al. (2025). We use its non-private configuration.

FLARE-FedAvg and FLARE-FedProx (ours) combine federated representation learning with real-instance neighborhoods, latent-locality weighting, support-aware completion, and weighted Ridge surrogates. FLARE-FedProx additionally applies proximal regularization during local encoder training.

Further details about these baselines can be found in Section 2.

4.3 Hyperparameters

We evaluate FLARE using Gaussian-kernel LIBSVM Chang & Lin (2011) and Deep Neural Network (DNN) LeCun et al. (2015) black-box models. Federated SCARF training uses 20 communication rounds, five local epochs, and full client participation. We evaluate FedAvg and FedProx (with $\mu = 0.01$). For local explanations, source data are partitioned into 15 K-medoids clusters; up to 200 relevant source instances and 150 target-training neighbors are used per query. All the hyperparameters are selected through grid search. Complete training, neighborhood, and reliability settings are provided in Appendix Table 4.

4.4 Evaluation Protocol

We report black-box predictive performance separately from explanation fidelity. Surrogate fidelity is measured using fidelity F1 Raza et al. (2025), Area under the curve (AUC), Matthews Correlation Coefficient (MCC) Matthews (1975), and balanced accuracy, while held-out fidelity uses stratified 80/20 neighborhood splits to assess local generalization. A neighborhood is *supported* when both predicted classes contain at least five instances and the raw and locality-weighted minority support are each at least $\tau_s = 0.10$. An explanation is *reliable* if it is supported and achieves held-out balanced accuracy ≥ 0.70 and MCC ≥ 0.50. Boundary fidelity is evaluated for queries with black-box probabilities in $[0.40, 0.60]$. Stability and robustness are measured using top-three-feature Jaccard consistency (JC) and the Local Lipschitz Estimate (LLE), respectively; higher JC and lower LLE indicate better performance. All reported results are averaged over five independent runs.

5 Experimental Evaluation and Discussion

We evaluate FLARE from three perspectives: comparison with federated explainability baselines, assessment of explanation reliability and robustness, and sensitivity analyses.

5.1 Comparison with Benchmark Baselines

Table 1 compares FLARE with FedRF-Surrogate, FL-LIME, and FastSHAP++ using fidelity F1, MCC, and AUC with 5 clients. These metrics measure how closely an explanation model reproduces its corresponding black-box predictions. Across all baselines, FLARE achieves the highest fidelity on every dataset and metric. FedRF-Surrogate is generally the strongest competing method but

Table 1: Fidelity comparison of FedRF-Surrogate, FL-LIME, FastSHAP++, and FLARE across seven classification tasks in the 5-client setting; ACS Income uses its predefined 50-client partition. Fidelity F1, MCC, and AUC are reported, with the best results shown in **bold**. For FLARE, FA and FP denote FedAvg and FedProx, while SVM and DNN identify the predictive model. FLARE achieves the highest fidelity across all datasets and metrics, with FedProx generally providing the strongest F1 and MCC.

Method	Adult			Bank			Student–STD1			Student–STD2			Diabetes–DTD1			Diabetes–DTD2			ACS Income		
	F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC
FedRF-Surrogate	0.674	0.596	0.936	0.536	0.460	0.873	0.891	0.729	0.966	0.869	0.677	0.970	0.627	0.648	0.958	0.267	0.340	0.941	0.414	0.400	0.898
FL-LIME	0.586	0.397	0.778	0.513	0.331	0.760	0.805	0.527	0.846	0.774	0.514	0.836	0.171	0.190	0.828	0.396	0.328	0.828	0.137	0.116	0.666
FastSHAP++	0.858	0.803	0.972	0.727	0.620	0.908	0.938	**0.916**	0.968	0.945	**0.912**	0.935	0.863	0.588	0.948	0.876	0.643	0.944	0.829	0.814	0.899
FLARE FA-SVM	0.806	0.728	0.967	0.661	0.654	0.973	0.954	0.889	0.992	0.966	0.899	**0.994**	0.252	0.251	0.835	0.273	0.268	0.796	0.745	0.650	0.910
FLARE FA-DNN	0.792	0.660	0.923	0.750	0.656	0.920	0.948	0.878	**0.993**	0.947	0.861	0.989	0.338	0.331	**0.997**	0.290	0.285	**0.998**	0.640	0.660	0.883
FLARE FP-SVM	**0.874**	**0.813**	**0.987**	**0.889**	**0.890**	**0.998**	**0.955**	0.887	0.992	**0.968**	0.905	0.993	0.818	0.821	**0.997**	0.891	0.887	0.997	**0.850**	**0.860**	**0.960**
FLARE FP-DNN	0.777	0.616	0.901	0.699	0.582	0.893	0.946	0.875	0.992	0.947	0.861	0.989	**0.893**	**0.880**	0.996	**0.900**	**0.888**	0.996	0.793	0.738	0.935

remains below the best FLARE configuration. The largest differences appear on Bank, Diabetes–DTD2, and ACS Income. On Diabetes–DTD2, for example, FLARE achieves F1 and MCC values of 0.900 and 0.888, compared with the best baseline values of 0.396 and 0.340, respectively.

FedProx generally produces the strongest F1 and MCC across all datasets. FLARE FP-SVM performs best on Adult, Bank, and ACS Income and achieves the highest F1 on both Student tasks. FLARE FP-DNN instead obtains the highest F1 and MCC on both Diabetes tasks. These results indicate that FLARE's explanation quality is not dependent on a single black-box architecture. Complete results for the 3-, 5-, and 10-client settings, including comparisons with all baselines, are reported in Appendix Table 8.

5.2 Reliability and Robustness

Table 2 evaluates the held-out fidelity, neighborhood support, reliability, stability, and robustness of FLARE-FedProx in the 5-client setting. Held-out fidelity is used here to determine whether surrogate agreement generalizes beyond the neighborhood observations used for fitting. SVM achieves higher held-out fidelity and reliable fractions on Adult, Bank, and both Student tasks. DNN performs better on Diabetes–DTD1 and achieves a slightly higher HO-F1 on Diabetes–DTD2.

DNN consistently achieves higher support rates, but greater support does not necessarily yield more reliable explanations. On Bank, for example, DNN achieves complete neighborhood support but a reliable fraction of only 0.070, whereas SVM achieves 0.680 support and 0.650 reliability. Thus, neighborhood support is necessary but insufficient; the surrogate must also generalize to held-out neighborhood observations. SVM further achieves higher JC and lower LLE across all six tasks, indicating more consistent feature selection and lower local sensitivity. Appendix Table 9 reports complete results for all FLARE variants in the 3-, 5-, and 10-client settings.

5.3 Sensitivity Analysis

To gain a deeper understanding of FLARE, we conduct some sensitivity experiments examining the contribution of its reliability-aware components and the effect of varying the number of clients.

5.3.1 Ablation Study

Table 3 evaluates FLARE after jointly disabling support-aware neighborhood completion and adaptive class-support weighting. Based on the results, the complete model improves F1, MCC, reliability, and AUC across all six available comparisons. The largest gains occur on Diabetes, where sufficient counter-class evidence is particularly important. On DTD1, removing both components reduces F1 from 0.818 to 0.189 and HO-MCC from 0.782 to 0.177. On DTD2, F1 decreases from 0.891 to 0.275 and HO-MCC from 0.855 to 0.246.

The gains on Adult, Bank, and both Student tasks are smaller but consistent, indicating that these mechanisms remain beneficial when the initial neighborhoods contain stronger class evidence. Because both components are disabled simultaneously, the experiment demonstrates their combined contribution but does not isolate their individual effects. Complete ablation results across all FLARE versions and client configurations are reported in Appendix Table 10.

Table 2: Reliability and robustness of FLARE-FedProx with 5 clients. Best values are shown in **bold**. Higher values are preferable for HO-F1, HO-MCC, Support, Reliable, and JC, whereas lower values are preferable for LLE. Overall, SVM provides stronger held-out fidelity, reliability, stability, and robustness on most datasets, while DNN consistently yields higher neighborhood support.

Metric	Adult		Bank		Student–STD1		Student–STD2		Diabetes–DTD1		Diabetes–DTD2	
	SVM	DNN	SVM	DNN	SVM	DNN	SVM	DNN	SVM	DNN	SVM	DNN
HO-F1	**0.810**	0.636	**0.863**	0.487	**0.927**	0.877	**0.941**	0.874	0.780	**0.838**	0.859	**0.863**
HO-MCC	**0.727**	0.403	**0.865**	0.280	**0.816**	0.719	**0.834**	0.691	0.782	**0.821**	**0.855**	0.851
Support	0.840	**0.960**	0.680	**1.000**	0.980	**1.000**	0.980	**1.000**	0.450	**0.620**	0.590	**0.630**
Reliable	**0.740**	0.200	**0.650**	0.070	**0.980**	0.930	**0.980**	0.940	0.420	**0.530**	**0.550**	**0.550**
JC	**0.741**	0.631	**0.966**	0.716	**0.945**	0.935	**0.977**	0.929	**0.835**	0.712	**0.812**	0.785
LLE	**14.116**	32.357	**12.310**	24.385	**1.360**	1.965	**1.583**	2.259	**2.153**	2.832	**3.153**	4.104

Table 3: Ablation study of FLARE across seven datasets in the 5-client setting; ACS Income uses its predefined 50-client partition. Full denotes the complete model, while w/o disables support-aware completion and adaptive class-support weighting. Across the available comparisons, Full FLARE improves all metrics, with the largest fidelity gains observed on the Diabetes tasks.

Metric	Adult		Bank		Student–STD1		Student–STD2		Diabetes–DTD1		Diabetes–DTD2		ACS Income	
	Full	w/o	Full	w/o	Full	w/o	Full	w/o	Full	w/o	Full	w/o	Full	w/o
F1	0.874	0.742	0.889	0.674	0.955	0.935	0.968	0.957	0.818	0.189	0.891	0.275	0.850	0.741
MCC	0.813	0.670	0.890	0.669	0.887	0.787	0.905	0.820	0.821	0.199	0.887	0.289	0.860	0.616
Reliable	0.740	0.450	0.650	0.430	0.980	0.750	0.980	0.790	0.420	0.080	0.550	0.120	0.810	0.56
AUC	0.987	0.987	0.998	0.979	0.992	0.990	0.993	0.992	0.997	0.961	0.997	0.936	0.960	0.943

5.3.2 Number of Clients

Appendix Tables 8 and 9 report FLARE results for data randomly distributed among three, five, and ten clients. For example, on Adult, FP-SVM remains stable as the number of clients increases: F1 values are 0.895, 0.874, and 0.876 for three, five, and ten clients, respectively. The corresponding MCC values are 0.820, 0.813, and 0.820, AUC remains between 0.987 and 0.989, and HO-F1 values are 0.816, 0.810, and 0.818. These limited variations indicate that increasing the number of client partitions does not materially reduce explanation fidelity on Adult. Similar patterns across the remaining datasets suggest that FLARE can operate under different federation sizes.

5.4 Non-IID Data Distribution

We evaluate FLARE under a naturally non-IID distribution using ACSIncome, where each of the 50 U.S. states constitutes one client and geographic differences induce heterogeneous local distributions. In this setting, FP-SVM achieves fidelity F1, MCC, and AUC values of 0.850, 0.860, and 0.960, respectively. These results demonstrate that FLARE maintains high explanation fidelity in a large federation with naturally heterogeneous clients.

6 Conclusion

We have presented **FLARE**, a federated framework for reliable local explanations under distributed and heterogeneous data. To the best of our knowledge, FLARE is the first framework to combine federated representation learning with client-local, real-instance neighborhood construction for query-specific surrogate explanations. Its reliability-aware design integrates federated latent locality, source–target neighborhoods, support-aware completion, and adaptive class-support weighting without exchanging raw data. Across five tabular datasets spanning seven classification tasks, FLARE consistently outperforms the evaluated federated explainability baselines in fidelity F1, MCC, and AUC, while held-out evaluation shows that its surrogates generalize within local neighborhoods. FLARE also remains effective across different client counts and achieves strong fidelity under the non-IID 50-state ACS Income federation. Future work will investigate formal privacy guarantees, communication efficiency, and extensions to additional data modalities. Overall, FLARE advances federated explainability from generating local explanations toward generating local explanations whose reliability can be explicitly assessed.

AI USE STATEMENT

Generative AI tools were used in the text refinement of the manuscript. AI-assisted suggestions were critically reviewed by the authors before submission.

ETHICS STATEMENT

This work studies local explainability in federated and geographically distributed data settings using existing benchmark datasets. No new human-subject data are collected, and the study does not involve direct interaction with individuals.

All datasets are used for research purposes under their established benchmark settings, and no attempt is made to identify individuals or recover private client records. The proposed method is intended as a tool for studying reliable local explanations in decentralized environments, not as a substitute for domain-specific validation, fairness assessment, legal review, or human oversight in high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce FLARE, including representation learning, neighborhood construction, surrogate fitting, reliability criteria, datasets, client configurations, black-box models, and evaluation metrics. Additional implementation details, hyperparameters, preprocessing steps, and extended results are reported in the appendix. Source code and experiment scripts are available as supplementary material and in the anonymous repository[2], with documented configurations, random seeds, and software dependencies.

REFERENCES

Azra Abtahi Fahliani, Amin Aminifar, and Amir Aminifar. Privacy-preserving federated interpretability. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, pp. 7592–7601. IEEE, 2024. doi: 10.1109/BigData62323.2024.10825590.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Kevin Bache and Moshe Lichman. Uci machine learning repository. 2013.

Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Transactions on Artificial Intelligence*, 5 (4):1429–1442, 2023.

Shuvo Biswas, Sajeeb Saha, Muhammad Shahin Uddin, and Rafid Mostafiz. An explainable and federated deep learning framework for skin cancer diagnosis. *PLOS ONE*, 20(7):e0324393, 2025.

Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez, David Sánchez, Adrian Flanagan, and Kuan Eeik Tan. Achieving security and privacy in federated learning systems: Survey, research challenges and future directions. *Engineering Applications of Artificial Intelligence*, 106:104468, 2021. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2021.104468. URL https://www.sciencedirect.com/science/article/pii/S095219762100316X.

Valerio Bonsignori, Luca Corbucci, Francesca Naretto, and Anna Monreale. Differentially private fastshap for federated learning model explainability. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Valerio Bonsignori, Luca Corbucci, Francesca Naretto, Riccardo Guidotti, and Anna Monreale. Interpreting federated learning by aggregating shap explanations. *IEEE Access*, 2026.

[2] https://anonymous.4open.science/r/FLARE-B858/README.md

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Priyam Choksi. 100,000-Diabetes Clinical Dataset, 2023. URL https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset. Accessed: 2025-03-07.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Pietro Ducange, Francesco Marcelloni, Giuseppe C. Miglionico, Andrea Renda, and Federico Ruffini. Federated SHAP: Privacy-Preserving and Consistent Post-hoc Explainability in Federated Learning. *Machine Learning*, 115(2):24, 2026a.

Pietro Ducange, Francesco Marcelloni, Giustino Claudio Miglionico, Alessandro Renda, and Fabrizio Ruffini. Federated shap: Privacy-preserving and consistent post-hoc explainability in federated learning. *Machine Learning*, 2026b. doi: 10.1007/s10994-025-06956-1.

Rami Haffar, David Sanchez, and Josep Domingo-Ferrer. Explaining predictions and attacks in federated learning via random forests. *Applied Intelligence*, 53(1):169–185, 2023.

Rajesh Kalakoti, Sven Nõmm, and Hayretdin Bahsi. Federated learning of explainable ai (fedxai) for deep learning-based intrusion detection in iot networks. *Computer Networks*, 270:111479, 2025. doi: 10.1016/j.comnet.2025.111479.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Melkamu Mersha, Khang Lam, Joseph Wood, Ali K. AlShami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.128111.

OpenML contributors. Student Mental Health dataset, 2020. URL https://www.openml.org/d/46753. Accessed: 2025-04-10.

Aryan Nikul Patel, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Ramalingam Murugan, Gokul Yenduri, and Thippa Reddy Gadekallu. A trustable federated learning framework for rapid fire smoke detection at the edge in smart home environments. *IEEE Internet of Things Journal*, 11(23):37708–37717, 2024.

Krishna Rani. FedLIME: Federated learning with LIME explanations. https://github.com/Krishna-Rani-t/Explainable_Federated_Learning, 2025. GitHub repository.

Rehan Raza, Guanjin Wang, Kok Wai Wong, Hamid Laga, and Marco Fisichella. ITL-LIME: Instance-Based Transfer Learning for Enhancing Local Explanations in Low-Resource Data Settings. In *ACM CIKM*, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Bukhoree Sahoh and Anant Choksuriwong. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7827–7843, June 2023. ISSN 1868-5145. doi: 10.1007/s12652-023-04594-w.

Zeren Tan, Yang Tian, and Jian Li. Glime: General, stable and local lime explanation. *Advances in neural information processing systems*, 36:36250–36277, 2023.

Titus Tunduny and Bernard Shibwabo. Explainable ai approaches in federated learning: Systematic review. *JMIR AI*, 5, 2026. ISSN 2817-1705. doi: https://doi.org/10.2196/69985. URL https://www.sciencedirect.com/science/article/pii/S2817170526000104.

A Appendix

Table 4: Main hyperparameters used in the FLARE experiments.

Parameter	Symbol	Value
Black-box models	–	G-LIBSVM, DNN
Federated strategies	–	FedAvg, FedProx
Learning rate	–	10^{-3}
Batch size	–	128
Communication rounds	T	20
Local epochs	E	5
FedProx coefficient	μ	0.1
Source clusters	M	15
Source instances	n_S	200
Target-training neighbors	K_T	150
Support threshold	τ_s	0.10
Queries per client	–	20

Table 5: Detailed experimental configuration used for FLARE.

Category	Setting	Value
Federated training	Communication rounds	20
	Local epochs	5
	Client configurations	3, 5, 10
	Aggregation algorithms	FedAvg, FedProx
	FedProx coefficient μ	0.01
Neighborhood	Source clusters K_C	15
	Source instances K_S	200
	Target neighbors K_T	150
Support-aware completion	Minimum minority fraction τ_s	0.10
	Maximum replacement fraction ρ	0.15
	Distance quantile q	0.95
	Distance expansion factor γ	1.25
Explanation	Neighborhood completion	Support-aware
	Surrogate class weighting	Adaptive
	Explanations per client	20
Black-box	Predictors	G-LIBSVM, DNN
	Black-box class weighting	None
Experiment control	Encoder mode	Federated
	Client partition	Balanced
	Random seed	42
	Maximum domain rows	0 (no restriction)

Table 6: Datasets used in the FLARE evaluation. Feature counts refer to original semantic features before one-hot encoding.

Dataset	N	#Feature	Task	Domain / client construction
Adult Bache & Lichman (2013)	48,842	14	Income $>$ \$50K	Local 75/25 source–target split.
Student–STD1 OpenML contributors (2020)	12,090	16	Depression	West $\rightarrow$ East (650 target instances).
Student–STD2 OpenML contributors (2020)	12,065	16	Depression	West $\rightarrow$ South (625 target instances).
Bank Bache & Lichman (2013)	45,211	15	Term deposit	Earlier 75% $\rightarrow$ later 25% observations.
ACS Income Ding et al. (2021)	State-dependent	10	Income $>$ \$50K	One U.S. state per client (50 natural clients).
Diabetes–DTD1 Choksi (2023)	97,501	10	Diabetes	Mainland U.S. $\rightarrow$ Virgin Islands (763 target instances).
Diabetes–DTD2 Choksi (2023)	97,261	10	Diabetes	Mainland U.S. $\rightarrow$ Guam (523 target instances).

Table 8: Fidelity comparison of FedRF-Surrogate, FL-LIME, FastSHAP++, and FLARE across seven classification tasks in the 5-client setting; ACS Income uses its predefined 50-client partition. F1, MCC, and AUC are reported, with the best results shown in **bold**. For FLARE, FA and FP denote FedAvg and FedProx, while SVM and DNN identify the predictive model. FLARE achieves the highest fidelity across all datasets and metrics, with FedProx generally providing the strongest F1 and MCC.

Dataset	C	FL-LIME			FedRF-Surrogate			FastSHAP++			FLARE				
		F1	MCC	AUC	F1	MCC	AUC	F1	MCC	AUC	**FL**	**Pred.**	F1	MCC	AUC
Adult	3	0.6245	0.4014	0.7731	0.6997	0.6157	0.9298	0.858	0.803	0.972	FA	SVM	0.7063	0.5527	0.9638
											FA	DNN	0.7246	0.5689	0.9038
											FP	SVM	**0.8947**	**0.8203**	0.9870
											FP	DNN	0.7583	0.5949	0.8888
											FA	SVM	0.8061	0.7284	0.9670
											FA	DNN	0.7916	0.6599	0.9231
	5	0.5863	0.3965	0.7775	0.6744	0.5956	0.9364	0.858	0.803	0.972	FP	SVM	0.8736	0.8128	0.9874
											FP	DNN	0.7765	0.6163	0.9012
											FA	SVM	0.8226	0.7533	0.9852
											FA	DNN	0.7849	0.6857	0.9419
											FP	SVM	0.8755	0.8200	**0.9889**
	10	0.4932	0.4480	0.8380	0.8191	0.7539	0.9621	0.858	0.803	0.972	FP	DNN	0.7951	0.6806	0.9305
Bank	3	0.4247	0.2894	0.7465	0.4886	0.4285	0.8573	0.727	0.620	0.908	FA	SVM	0.7339	0.7174	0.9407
											FA	DNN	0.7461	0.6631	0.9305
											FP	SVM	**0.9377**	**0.9331**	**0.9992**
											FP	DNN	0.6902	0.5617	0.8810
											FA	SVM	0.6609	0.6544	0.9725
											FA	DNN	0.7498	0.6556	0.9198
	5	0.5131	0.3308	0.7599	0.5358	0.4597	0.8727	0.727	0.620	0.908	FP	SVM	0.8892	0.8896	0.9979
											FP	DNN	0.6992	0.5821	0.8928
											FA	SVM	0.7811	0.7743	0.9960
											FA	DNN	0.7451	0.6448	0.9153
											FP	SVM	0.9227	0.9090	0.9967
	10	0.5750	0.4225	0.8237	0.7816	0.6721	0.9280	0.727	0.620	0.908	FP	DNN	0.7107	0.5988	0.9001
Student–STD1	3	0.8152	0.4759	0.8215	0.9126	0.7806	0.9630	0.938	0.916	0.968	FA	SVM	0.9145	0.8279	0.9907
											FA	DNN	0.9274	0.8368	0.9841
											FP	SVM	0.9427	0.8742	0.9916
											FP	DNN	0.9181	0.8285	0.9851
											FA	SVM	0.9541	0.8894	0.9922
											FA	DNN	0.9476	0.8780	0.9925
	5	0.8049	0.5273	0.8459	0.8914	0.7287	0.9663	0.938	**0.916**	0.968	FP	SVM	0.9552	0.8872	0.9915
											FP	DNN	0.9463	0.8745	0.9917
											FA	SVM	0.9660	0.9049	0.9943
											FA	DNN	0.9655	0.9042	0.9946
											FP	SVM	**0.9668**	0.9067	0.9943
	10	0.7920	0.5264	0.8465	0.9106	0.7430	0.9544	0.938	0.916	0.968	FP	DNN	0.9660	0.9057	**0.9947**
Student–STD2	3	0.7934	0.5010	0.8307	0.9009	0.6360	0.9032	0.945	0.912	0.935	FA	SVM	0.9422	0.8537	**0.9941**
											FA	DNN	0.9320	0.8487	0.9854
											FP	SVM	0.9574	0.8985	0.9935
											FP	DNN	0.9289	0.8383	0.9856
											FA	SVM	0.9663	0.8993	0.9935
											FA	DNN	0.9466	0.8612	0.9893
	5	0.7739	0.5141	0.8361	0.8690	0.6772	0.9698	0.945	**0.912**	0.935	FP	SVM	**0.9676**	0.9051	0.9934
											FP	DNN	0.9473	0.8609	0.9885
											FA	SVM	0.9515	0.8781	0.9893
											FA	DNN	0.9559	0.8940	0.9935
											FP	SVM	0.9447	0.8720	0.9892
	10	0.7893	0.5435	0.8540	0.9205	0.7789	0.9652	0.945	0.912	0.935	FP	DNN	0.9494	0.8876	0.9931
Diabetes–DTD1	3	0.0821	0.1269	0.8560	0.7963	0.8127	0.9880	0.863	0.588	0.948	FA	SVM	0.2217	0.1774	0.8440
											FA	DNN	0.3412	0.2671	0.9908
											FP	SVM	**0.9265**	**0.9123**	0.9951
											FP	DNN	0.9109	0.8935	0.9943
											FA	SVM	0.2518	0.2506	0.8346
											FA	DNN	0.3384	0.3310	0.9970
	5	0.1709	0.1895	0.8280	0.6267	0.6478	0.9576	0.863	0.588	0.948	FP	SVM	0.8179	0.8210	0.9965
											FP	DNN	0.8934	0.8796	0.9959
											FA	SVM	0.2024	0.1978	0.9004
											FA	DNN	0.3690	0.3679	**0.9982**
											FP	SVM	0.6583	0.6714	0.9978
	10	0.0967	0.1189	0.8306	0.4800	0.6139	0.9546	0.863	0.588	0.948	FP	DNN	0.8923	0.8857	0.9970
Diabetes–DTD2	3	0.0243	0.0424	0.8255	0.8000	0.8258	0.9919	0.876	0.643	0.944	FA	SVM	0.2381	0.1994	0.7565
											FA	DNN	0.2471	0.2279	**0.9982**
											FP	SVM	0.8993	**0.8907**	0.9960
											FP	DNN	0.8978	0.8793	0.9942
											FA	SVM	0.2726	0.2676	0.7962
											FA	DNN	0.2904	0.2846	0.9976
	5	0.396	0.328	0.828	0.2667	0.3395	0.9409	0.876	0.643	0.944	FP	SVM	0.8912	0.8866	0.9965
											FP	DNN	**0.8996**	0.8884	0.9956
											FA	SVM	0.3322	0.3446	0.9391
											FA	DNN	0.4725	0.4691	0.9969
											FP	SVM	0.7951	0.8035	0.9978
	10	0.1538	0.1824	0.8161	0.5167	0.6500	0.9948	0.876	0.643	0.944	FP	DNN	0.8973	0.8859	0.9969
ACS Income	50	0.1366	0.1164	0.6657	0.4143	0.4005	0.8977	0.829	0.814	0.899	FA	SVM	0.54	0.56	0.84
											FP	SVM	**0.85**	**0.86**	**0.96**

Table 9: Reliability and robustness of FLARE-FedProx with 5 clients. Best values are shown in **bold**. Higher values are preferable for HO-F1, HO-MCC, Support, Reliable, and JC, whereas lower values are preferable for LLE. Overall, SVM provides stronger held-out fidelity, reliability, stability, and robustness on most datasets, while DNN consistently yields higher neighborhood support.

Dataset	C	FL	Pred.	HO-F1	HO-MCC	Support	Reliable	JC	LLE
Adult	3	FA	SVM	0.5219	0.2801	0.5167	0.1500	0.7080	24.4672
		FA	DNN	0.4292	0.1745	0.8333	0.0667	0.7655	44.9300
		FP	SVM	0.8162	0.7033	0.8833	0.6667	0.6815	23.7663
		FP	DNN	0.6033	0.3627	0.9000	0.0833	0.6257	31.9342
	5	FA	SVM	0.5836	0.4529	0.7500	0.3400	0.7697	25.4543
		FA	DNN	0.5167	0.2626	0.8800	0.1600	0.6729	38.7161
		FP	SVM	0.8102	0.7266	0.8400	**0.7400**	0.7405	14.1164
		FP	DNN	0.6356	0.4033	**0.9600**	0.2000	0.6308	32.3565
	10	FA	SVM	0.7311	0.6141	0.6900	0.5600	**0.8348**	15.0808
		FA	DNN	0.5556	0.3770	0.7700	0.1700	0.8076	16.8371
		FP	SVM	**0.8176**	**0.7307**	0.8300	0.7100	0.8325	**14.0278**
		FP	DNN	0.6495	0.4697	0.8200	0.2450	0.7699	15.4896
Bank	3	FA	SVM	0.6195	0.5769	0.6500	0.5667	0.9323	18.5169
		FA	DNN	0.3678	0.1784	0.8667	0.0833	0.8433	32.4465
		FP	SVM	**0.9081**	**0.9058**	0.7667	**0.7667**	**0.9733**	13.6473
		FP	DNN	0.5014	0.2830	**1.0000**	0.0500	0.6362	33.5627
	5	FA	SVM	0.5745	0.5536	0.5900	0.5300	0.9314	15.5746
		FA	DNN	0.3988	0.1868	0.9200	0.1000	0.8127	24.1810
		FP	SVM	0.8631	0.8650	0.6800	0.6500	0.9655	12.3101
		FP	DNN	0.4867	0.2795	**1.0000**	0.0700	0.7162	24.3851
	10	FA	SVM	0.6957	0.6781	0.5800	0.5700	0.9573	10.5302
		FA	DNN	0.4441	0.2402	0.9700	0.0750	0.8612	19.9635
		FP	SVM	0.8941	0.8773	0.6400	0.6350	0.9643	**8.3507**
		FP	DNN	0.4873	0.2897	**1.0000**	0.0700	0.8224	19.6452
Student–STD1	3	FA	SVM	0.8335	0.6878	0.9000	0.7833	0.8345	3.0582
		FA	DNN	0.7862	0.5784	0.9833	0.5833	0.8103	2.2758
		FP	SVM	0.9169	0.8247	0.9833	**0.9833**	0.9032	1.7552
		FP	DNN	0.8369	0.6557	**1.0000**	0.8000	0.7845	1.5726
	5	FA	SVM	0.9103	0.7912	0.9700	0.9500	0.9425	1.4679
		FA	DNN	0.8613	0.6922	0.9900	0.8700	0.9256	2.6231
		FP	SVM	0.9273	0.8162	0.9800	0.9800	0.9445	1.3601
		FP	DNN	0.8772	0.7194	**1.0000**	0.9300	0.9351	1.9649
	10	FA	SVM	0.9388	0.8262	0.9846	0.9774	0.9692	1.0539
		FA	DNN	0.8946	0.7206	**1.0000**	0.9248	0.9835	1.6445
		FP	SVM	**0.9405**	**0.8311**	0.9846	0.9699	0.9534	**0.9928**
		FP	DNN	0.8946	0.7207	**1.0000**	0.9398	**0.9850**	1.6072
Student–STD2	3	FA	SVM	0.8677	0.6914	0.9167	0.7500	0.8567	2.9401
		FA	DNN	0.7859	0.5565	0.9667	0.5333	0.6993	3.8379
		FP	SVM	0.9322	0.8282	0.9667	0.9500	0.8967	1.8213
		FP	DNN	0.8413	0.6455	0.9833	0.8500	0.7845	2.6515
	5	FA	SVM	0.9312	0.8079	0.9600	0.9600	0.9608	1.8565
		FA	DNN	0.8509	0.6480	**1.0000**	0.8400	0.9399	2.7153
		FP	SVM	**0.9411**	**0.8335**	0.9800	**0.9800**	**0.9770**	1.5834
		FP	DNN	0.8743	0.6912	**1.0000**	0.9400	0.9293	2.2588
	10	FA	SVM	0.9134	0.7744	**1.0000**	0.9769	0.9423	**1.1553**
		FA	DNN	0.8680	0.6815	**1.0000**	0.9231	0.9762	2.2291
		FP	SVM	0.9090	0.7690	**1.0000**	0.9769	0.9454	1.1920
		FP	DNN	0.8644	0.6769	**1.0000**	0.9154	0.9762	2.0990
Diabetes–DTD1	3	FA	SVM	0.1498	0.1051	0.3500	0.0167	0.5498	14.2198
		FA	DNN	0.2043	0.1255	0.2667	0.0167	0.6533	7.2828
		FP	SVM	**0.9102**	**0.8961**	0.8000	**0.7667**	0.8055	2.4381
		FP	DNN	0.8750	0.8518	**0.8167**	0.6667	0.7338	2.5612
	5	FA	SVM	0.2073	0.1950	0.2400	0.1500	0.5588	7.3815
		FA	DNN	0.2412	0.2116	0.2700	0.1000	0.7044	6.0506
		FP	SVM	0.7796	0.7818	0.4500	0.4200	0.8350	2.1533
		FP	DNN	0.8381	0.8214	0.6200	0.5300	0.7120	2.8315
	10	FA	SVM	0.1659	0.1665	0.1125	0.0875	0.6982	3.3458
		FA	DNN	0.3130	0.3055	0.2375	0.1812	0.7804	3.2451
		FP	SVM	0.6194	0.6324	0.2000	0.1938	**0.8602**	**1.8983**
		FP	DNN	0.8504	0.8434	0.4375	0.4125	0.8245	2.4756
Diabetes–DTD2	3	FA	SVM	0.1368	0.0944	0.1667	0.0167	0.4818	12.3073
		FA	DNN	0.1906	0.1388	0.2833	0.0833	0.7220	8.4387
		FP	SVM	0.8633	0.8570	**0.7333**	**0.6667**	0.8110	3.1353
		FP	DNN	0.8590	0.8355	**0.7333**	0.6167	0.7490	3.3054
	5	FA	SVM	0.1943	0.1839	0.2600	0.0800	0.6052	9.5595
		FA	DNN	0.1916	0.1611	0.2900	0.0800	0.7163	7.9459
		FP	SVM	0.8585	0.8551	0.5900	0.5500	**0.8121**	3.1529
		FP	DNN	0.8632	0.8514	0.6300	0.5500	0.7849	4.1039
	10	FA	SVM	0.2929	0.3017	0.1364	0.0909	0.6850	3.7504
		FA	DNN	0.4306	0.4276	0.3273	0.2727	0.7995	3.6128
		FP	SVM	0.7666	0.7776	0.2273	0.2091	0.8005	**2.3346**
		FP	DNN	**0.8706**	**0.8610**	0.5455	0.4727	0.7801	2.9498

Table 10: Ablation results without adaptive surrogate class weighting and without support-aware neighborhood completion. All runs use FedProx (FP). Best values per dataset are shown in **bold**. Higher values are better for all reported metrics.

Dataset	C	FL	Pred.	Fidelity F1	Fidelity MCC	HO-MCC	Reliable	AUC
Adult	3	FP	SVM	**0.7556**	0.6463	0.5918	0.4167	0.9734
		FP	DNN	0.5481	0.4030	0.2289	0.0500	0.8847
	5	FP	SVM	0.7421	**0.6703**	0.5978	0.4500	**0.9866**
		FP	DNN	0.5873	0.4454	0.2725	0.0800	0.8979
	10	FP	SVM	0.7326	0.6609	**0.6005**	**0.4900**	0.9780
		FP	DNN	0.5313	0.4297	0.2657	0.1100	0.9275
Bank	3	FP	SVM	**0.6969**	**0.6965**	**0.8750**	**0.5167**	**0.9933**
		FP	DNN	0.5141	0.4459	0.2010	0.0333	0.8753
	5	FP	SVM	0.6738	0.6687	0.7987	0.4300	0.9788
		FP	DNN	0.5407	0.4822	0.1992	0.0300	0.8828
	10	FP	SVM	0.6728	0.6663	0.7544	0.3600	0.9925
		FP	DNN	0.5486	0.4872	0.2252	0.0750	0.8911
Student–STD1	3	FP	SVM	0.9145	0.7722	0.7027	0.7500	0.9900
		FP	DNN	0.8900	0.7385	0.5823	0.6167	0.9836
	5	FP	SVM	0.9349	0.7865	0.7141	0.7500	0.9902
		FP	DNN	0.9256	0.7976	0.6289	**0.7800**	0.9907
	10	FP	SVM	**0.9607**	0.8414	**0.7421**	0.7594	0.9914
		FP	DNN	0.9544	**0.8485**	0.6515	0.7594	**0.9936**
Student–STD2	3	FP	SVM	0.9411	0.7642	0.7053	0.7500	0.9921
		FP	DNN	0.9093	0.7050	0.5400	0.6833	0.9827
	5	FP	SVM	**0.9571**	0.8200	**0.7657**	0.7900	0.9916
		FP	DNN	0.9433	0.7999	0.6178	0.7400	0.9869
	10	FP	SVM	0.9355	0.8150	0.7114	0.7615	0.9864
		FP	DNN	0.9407	**0.8360**	0.6217	**0.8000**	**0.9924**
Diabetes–DTD1	3	FP	SVM	0.2387	0.2491	**0.3091**	0.1500	0.9809
		FP	DNN	**0.3202**	**0.3223**	0.2616	**0.1833**	0.9897
	5	FP	SVM	0.1892	0.1985	0.1771	0.0800	0.9608
		FP	DNN	0.2504	0.2571	0.2081	0.1400	0.9925
	10	FP	SVM	0.1151	0.1204	0.1049	0.0125	0.9614
		FP	DNN	0.1508	0.1602	0.1230	0.0563	**0.9950**
Diabetes–DTD2	3	FP	SVM	0.2746	0.2849	0.2187	0.0667	0.9130
		FP	DNN	0.2170	0.2241	0.1801	**0.1500**	0.9873
	5	FP	SVM	**0.2750**	**0.2888**	0.2459	0.1200	0.9361
		FP	DNN	0.2129	0.2212	0.1795	0.1300	0.9924
	10	FP	SVM	0.2091	0.2310	**0.2884**	0.0091	0.9452
		FP	DNN	0.2233	0.2307	0.2405	0.0818	**0.9946**