iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82269

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

SSE-AGG: Shift-Stable Explanations via Aggregation for Federated Learning in Healthcare

Prof. MPS Bhatia[1], Aaryan Mahajan[2], Harsh Rawat[3], Shreya Singh[4]
Department of Computer Science and Engineering, Netaji Subhas University of Technology

Abstract: ***Clinical machine learning systems that behave inconsistently across hospitals are hard to trust, and harder still to govern. This paper takes the position that explanation stability under distribution shift is not a secondary concern—it is a core requirement for federated learning in healthcare. We present Shift-Stable Explanations via Aggregation (SSE-AGG), a server-side framework that makes explanation consistency a first-class objective during training. Rather than treating all client updates equally, SSE-AGG clusters clients based on their model updates and then steers aggregation toward those whose feature-importance sketches—computed via permutation importance over a small shared reference set—agree with the cohort consensus. The method requires no data sharing, no additional libraries, and no changes to client-side code, making it straightforward to layer on top of existing FL pipelines. It works with FedAvg, FedProx, and SCAFFOLD. We define and measure three explanation-stability metrics: pairwise L_1 drift across clients, consensus drift across rounds, and top-k Jaccard overlap. Experiments on controlled non-IID partitions show that SSE-AGG meaningfully reduces explanation drift—especially under high heterogeneity—without degrading predictive accuracy. Taken together, the results suggest that explanation-aware aggregation is both feasible and practically useful as a step toward more interpretable and governable federated clinical systems.***
Index Terms: ***Federated Learning, Explanation Stability, Non-IID Data, Model Interpretability, Distribution Shift, Trustworthy AI***

I. INTRODUCTION

A recurring frustration in clinical ML deployment is that models trained at one hospital often perform quite differently at another. Patient populations differ, documentation habits differ, care protocols differ—and all of these factors introduce distributional shift that quietly erodes both predictive reliability and the trustworthiness of risk estimates. Federated learning (FL) is widely seen as the right structural response to this problem: it lets hospitals collaborate on model training without anyone having to hand over their patient records. But FL is not a free lunch. When client data is non-IID, local updates can pull the global model in conflicting directions, and the resulting model may generalize well on average while failing badly at the worst-performing site. In healthcare, worst-case failure is the scenario that matters most.

There is a second problem that tends to get less attention but is equally important in practice: explanation stability. Even when a federated model achieves acceptable aggregate accuracy, different hospitals may encounter meaningfully different rationales for similar patient profiles. One site might find that age and creatinine dominate the model's risk estimates; another might see oxygen saturation and medication history weighted far more heavily. When the same model produces such divergent explanations across sites—or shifts its explanations over training rounds—it becomes genuinely difficult for clinicians to build intuition about the system, and it becomes nearly impossible for governance teams to audit it consistently. The question of which clinical signals are driving predictions is not just a nice-to-have—it underlies the trust that is necessary for adoption.

Our work addresses both dimensions. We develop an explanation-aware aggregation strategy that operates entirely on the server side, using lightweight sketches of feature importance to quantify how much clients agree in their learned rationales. Clients that deviate substantially from the consensus receive lower aggregation weight, while those that align with the cohort are trusted more. We combine this with objectives targeting worst-site performance and calibration, and we track explanation drift over training rounds as a practical monitoring signal.

We evaluate the resulting framework on heterogeneous client partitions derived from both tabular clinical data and image-based skin-lesion classification. Our primary claim is not that SSE-AGG is optimal across all possible settings—it is that explanation stability is a measurable, controllable property of a federated system, and that a modest intervention at the aggregation layer is sufficient to improve it substantially without sacrificing predictive utility.

II. RELATED WORK

A. Explanation Robustness Under Distribution Shift

Li et al. [1] studied whether model explanations hold up under distribution shift and proposed Distributionally Robust Explanations (DRE), which enforce cross-distribution consistency during training. The results are convincing within a centralized setting, but the optimization is complex and the mixup-based assumptions do not translate easily into the federated case. Lasko et al. [2] took a more diagnostic approach, asking why probabilistic clinical models fail to transport between sites. They point to workflow differences, prevalence changes, overfitting, and data leakage as the primary culprits, and argue that leave-one-site-out evaluation is essential for detecting these issues before deployment. Avati et al. [3] introduced BEDS-Bench, a benchmark specifically designed to quantify out-of-distribution degradation in EHR models, and documented large performance drops under realistic distribution shifts across MIMIC-III and PICDB datasets.

B. Federated Optimization Under Heterogeneity

The foundational methods for handling non-IID data in FL are FedProx [5] and SCAFFOLD [6]. FedProx adds a proximal regularization term to the local objective so that client updates do not stray too far from the global model. SCAFFOLD addresses client drift more directly using control variates that correct the gradient estimates on each client. Both methods improve convergence under heterogeneity, but neither considers the consistency of learned explanations as part of the optimization. SSE-AGG is designed to complement both.

C. Client Clustering and Cohort-Aware Aggregation

Sattler et al. [7] proposed grouping clients by similarity and training separate per-cluster models, which works well when the data partitions are truly distinct. Ghosh et al. [8] introduced IFCA, which dynamically discovers client clusters during training and enables personalized models within each group. Our hierarchical aggregation scheme borrows from this line of work, but applies clustering at the level of model update vectors rather than data labels, and uses it to structure explanation-based reweighting within cohorts.

D. Robust Aggregation

Several papers have explored using external signals to modulate the influence of individual clients during aggregation. FLTrust [4] maintains a small trusted dataset on the server and uses it to score incoming client updates against Byzantine noise. FoolsGold [9] reduces the weight of clients whose updates are suspiciously similar, targeting Sybil attacks. Krum [10] selects client updates by proximity in parameter space. These approaches provide useful precedent for the idea that aggregation weights need not be fixed—they can reflect how much a given client should be trusted in a particular round.

E. Feature Importance Methods

The explanation sketches in SSE-AGG are built on permutation importance [11], which estimates feature relevance by measuring the drop in model performance when each feature is permuted independently. Fisher et al. [12] extended this idea to Model Class Reliance, quantifying importance across an entire equivalence class of models. Both methods are model-agnostic and require no gradient access, which makes them well-suited for a federated setting where clients may use different local implementations.

III. PROBLEM STATEMENT

We consider a standard federated learning setup with K clients training over T rounds. Each client k holds local data $\ddot{Y}_k = \{(x^i k, y^i k)\}_{nk}$ drawn from a distribution P_k. Distributions differ across clients—sometimes substantially—and this non-IID structure is the central source of difficulty. At round t, clients perform local updates and return their parameters to the server, which aggregates them into a global model θ^t.

Standard aggregation schemes—FedAvg, FedProx, SCAFFOLD—optimize for predictive accuracy and do not say anything about the stability of model explanations. To make this concrete, let $E^t_k \in \mathbb{R}^d$ denote the feature-importance vector produced by client k's model at round t. We care about two distinct failure modes:

1) Cross-client drift: $\frac{2}{k(k-1)} \Sigma_{i<j} \|E_i^t - E_j^t\|_1$ is large, meaning clients are relying on different features to make the same predictions at the same round.

2) Round-to-round drift: $\bar{E}^t - \bar{E}^{t-1}\|_1$ is large, meaning the global consensus on feature relevance is shifting unpredictably as training progresses.

We want an aggregation rule that keeps both of these quantities small, while keeping the predictive quality of the global model at least as good as standard baselines.

IV. METHODOLOGY

A. *Data and Notation*

Let $D = \{(x_i, y_i)\}_{i=1}^N$ with $y_i \in \{0,1,2\}$ denote the full dataset. After NBERT mean-pooled embedding, each sample is represented as $x_i \in \mathbb{R}^d$. Data is partitioned across K clients via Dirichlet allocation $D_k \sim$ Dirichlet(α), where smaller α produces more extreme label skew. Features are standardized as $\tilde{x} = (x - \mu)/(\sigma + \varepsilon)$ before training.

B. *Local Model and Training Objective*

Each client trains a two-layer MLP with hidden activations $H = \text{ReLU}(ZW_1 + b_1)$, output logits $\hat{Y} = HW_2 + b_2$, and class probabilities $p_i = \text{softmax}(\hat{y}_i)$. The local objective is cross-entropy with ℓ_2 regularization:

$$\mathcal{L}_k(\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i^T log p_i + \frac{\lambda_{reg}}{2}(||W_1||_2^2 + || W_2||_2^2) \quad (1)$$

When running with FedProx, a proximal term $(\mu/2)\|\theta - \theta^t\|_2$ is added to prevent excessive deviation from the global model. Parameters are updated via SGD: $\theta_k \leftarrow \theta_k - \eta\nabla\mathcal{L}_k(\theta_k)$ over E local epochs.

C. *Explanation Sketches via Permutation Importance*

At each round, every client computes a compact feature-importance vector using the server-held reference set $R \in \mathbb{R}^{m\times d}$ (Algorithm 3). The idea is simple: score the model on R, then measure how much performance drops when each feature is permuted. A baseline score $A_0 = \text{SCORE}(f\theta; R)$ is recorded first. For feature j, we permute that column and recompute:

$$PI_j = \max\{0, A_0 - \text{SCORE}(f_\theta ; \pi_j(R))\} \quad (2)$$

The raw importances are then normalized onto the probability simplex:

$$E_{k,j}^t = \frac{PI_j}{\sum_{j'=1}^{d} PI_{j'} + \varepsilon} \quad (3)$$

If communication efficiency matters, we retain only the top-q entries and renormalize, reducing each sketch to roughly 2q floats.

D. *Robust SSE-AGG: Four-Component Aggregation*

The server aggregates client updates using a weighted combination of four signals (Algorithm 1). Each component targets a different failure mode.

Data-size weights are the standard FedAvg baseline:

$$w_k^{data} = \frac{n_k}{\sum_{j=1}^{k} n_j} \quad (4)$$

Explanation-agreement weights give more influence to clients whose sketches agree with the round consensus $\bar{E}^t = (1/K)\Sigma_k E^t_k$:

$$c_k^t = \frac{1}{\varepsilon + \|E_k^t - \bar{E}^t\|_1}, \quad w_k^{expl} = \frac{c_k^t}{\sum_{j=1}^{k} c_j^t} \quad (5)$$

Worst-site weights up-weight clients with higher local validation loss, nudging the global model toward better worst-case generalization :

$$w_k^{worst} = \frac{exp(\beta l_k^t)}{\sum_{j=1}^{k} exp(\beta l_j^t)} \quad (6)$$

Calibration weights similarly up-weight clients with higher Expected Calibration Error, encouraging better-calibrated aggregation:

$$w_k^{cal} = \frac{exp(\gamma \cdot ECE_k^t)}{\sum_{j=1}^{k} exp(\gamma \cdot ECE_j^t)} \quad (7)$$

These four signals are blended with non-negative coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$, $\Sigma_i \lambda_i = 1$:

$$w_k^{agg} = \lambda_1 w^{data} + \lambda_2 w^{expl} + \lambda_3 w^{worst} + \lambda_4 w^{cal} \quad (8)$$

Setting $\lambda_1 = 1$ exactly recovers FedAvg. The global update is:

$$\theta^{t+1} = \sum_{k=1}^{K} w_k^{agg} \theta_k^{t+1} \quad (9)$$

E. Cohort-Based Hierarchical Aggregation

When client populations are legitimately heterogeneous—pediatric vs. adult, or primary care vs. ICU—it is not appropriate to penalize clients for disagreeing with a single global consensus. In these cases, we first cluster clients into *m* cohorts via k-means on update vectors $\Delta^t_k = \theta^{t+1}_k - \theta^t$, then apply SSE-AGG within each cohort separately:

$$\hat{w}_{k,c}^{agg} = \frac{w_k^{agg}}{\sum_{j \in C_c} w_j^{agg}}, \quad \theta_c^{t+1} = \sum_{k \in C_c} \hat{w}_{k,c}^{agg} \theta_k^{t+1} \quad (10)$$

Cohort-level models are then combined with cohort-size weights:

$$\theta_k^{t+1} = \sum_{c=1}^{m} \frac{|C_c|}{\sum_{j=1}^{m} |C_j|} \theta_c^{t+1} \quad (11)$$

This allows explanation stability to be enforced within clinically coherent groups without suppressing genuine variation between them.

F. Post-hoc Temperature Scaling

After training, we calibrate the global model using temperature scaling (Algorithm 4). A scalar $T^\star > 0$ is chosen to minimize negative log-likelihood on a held-out validation set:

$$T^\star = \underset{T>0}{\arg min} \left(-\sum_{(x_i, y_i) \in v} \log softmax\left(\frac{\hat{y}_i}{T}\right)_{y_i} \right) \quad (12)$$

At inference, predictions use $\text{softmax}(\frac{\hat{y}}{T^\star})$. Temperature scaling does not affect rank-order predictions, so it can be applied without any retraining.

G. Algorithms

Algorithm 1: Robust SSE-AGG with Cohorts, Explanation Sketches, and Temperature Scaling

Input: Rounds T; clients $k \in \{1,\ldots,K\}$; local epochs E; lr η; reference set R; blend weights $\lambda_1,\lambda_2,\lambda_3,\lambda_4 \geq 0$, $\Sigma\lambda_i = 1$; temperatures $\beta,\gamma > 0$; $\varepsilon > 0$; optional top-q sparsity; optional cohort count m.

Output: Global parameters θ^T; calibrated temperature $T^\star$ (optional).

Server: Initialize θ^0.

for t = 0 to T−1 do

  Broadcast θ^t and R to all clients.

  for k = 1 to K do [in parallel]

    $(\theta^{t+1}_k, \Delta^t_k, E^t_k, l^t_k, ECE^t_k, n_k) \leftarrow ClientUpdate(k, \theta^t, R, E, \eta, \varepsilon, q)$

  $\bar{E}^t \leftarrow \frac{1}{K} \sum_k E^t_k$ // consensus sketch

  for k: $c^t_k \leftarrow (\varepsilon + \|E^t_k - \bar{E}^t\|_1)^{-1}$; $w_k^{expl} \leftarrow \frac{c^t_k}{\sum_j c^t_j}$

  for k: $w^d_k \leftarrow \frac{n_k}{\sum_j n_j}$

  for k: $w_k^{worst} \leftarrow \frac{e^{\beta l^t_k}}{\sum_j \beta l^t_j}$

  for k: $w_k^{cal} \leftarrow \frac{e^{\gamma ECE^t_k}}{\sum_j \gamma ECE^t_j}$

  for k: $w_k^{agg} \leftarrow \lambda_1 w_k^{data} + \lambda_2 w_k^{expl} + \lambda_3 w_k^{worst} + \lambda_4 w_k^{cal}$

if m specified then
$\{C_c\} \leftarrow$ KMeans($\{\Delta_k^t\}$, m)
for c = 1 to m do
$\widehat{w}_{k,c} \leftarrow \frac{w_k^{agg}}{\sum_{k \in C_c} w_j^{agg}}$ for k $\in C_c$;
$\theta_c^{t+1} \leftarrow \sum_{k \in C_c} \widehat{w}_{k,c} \theta_k^{t+1}$
$\theta^{t+1} \leftarrow \sum_c \frac{|C_c|}{\sum_j |C_j|} \quad \theta_c^{t+1}$
else
$\theta^{t+1} \leftarrow \sum_k w_k^{agg} \theta_k^{t+1}$
$T\star \leftarrow$ TemperatureScaling(θ^T)
return θ^T, $T\star$.

Algorithm 2: ClientUpdate(k, θ^t, R, E, η, ε, q)
Input: Client k; global params θ^t; reference R; epochs E; lr η; ε; sparsity q.
Output: θ_k^{t+1}, Δ_k^{t+1}, sketch E_k^t, loss L_k^t, ECE_k^t, n_k.

$\theta_k \leftarrow \theta^t$
for e = 1 to E do
$\theta_k \leftarrow \theta_k - \eta \nabla \mathcal{L}_k(\theta_k)$
$\theta_k^{t+1} \leftarrow \theta_k$; $\Delta_k^t \leftarrow \theta_k^{t+1} - \theta^t$; $n_k \leftarrow |\mathcal{D}_k|$
$E_k^t \leftarrow$ SKETCHPERMIMP(θ_k^{t+1}, R, ε, q)
$l_k^t \leftarrow$ ValLoss(θ_k^{t+1}); $ECE_k^t \leftarrow$ ECE(θ_k^{t+1})
return (θ_k^{t+1}, Δ_k^t, E_k^t, l_k^t, ECE_k^t, n_k)

Algorithm 3: SKETCHPERMIMP(θ, R, ε, q)
Input: Model θ; reference R; ε > 0; optional top-q.
Output: Sketch E $\in \mathbb{R}^d$, $\sum_j E_j = 1$.

$A_0 \leftarrow$ SCORE(f_θ; R)
for j = 1 to d do
$A_j \leftarrow$ SCORE(f_θ; $\pi_j(R)$); $PI_j \leftarrow \max\{0, A_0 - A_j\}$
for j: $E_j \leftarrow \frac{PI_j}{\sum_{j'} PI_{j'} + \epsilon}$
if q specified then
E $\leftarrow$ TOPQ(E, q); E $\leftarrow \frac{E}{\sum_j E_j + \epsilon}$
return E

Algorithm 4: TEMPERATURESCALING(θ)
Input: Trained θ; labeled validation set Ÿ.
Output: Optimal temperature $T\star > 0$.

$T\star \leftarrow argmax_{T>0} \left(- \sum_{(x_i, y_i) \in v} \log softmax(\frac{\hat{y}_i}{T})_{y_i} \right)$
return $T\star$

V. EVALUATION METRICS

We evaluate both predictive quality and explanation stability. Our stability metrics are defined so that lower values always mean more consistent explanations, and higher Jaccard means more agreement on which features matter most.

Cross-client explanation drift (lower is better):

$$Drift^t_{pair} = \frac{2}{K(K-1)} \sum_{i<j} \left\| E^t_i - E^t_j \right\| \quad (13)$$

Round-to-round consensus drift (lower is better):

$$Drift^t_{round} = \| \bar{E}^t - \bar{E}^{t-1} \|_1 \quad (14)$$

Top-k Jaccard similarity (higher is better): Let S^i_k = top-k($\tilde{e}_i$) denote the top-k feature indices for client i:

$$J_{ij} = \frac{\left| s^i_k \cap s^j_k \right|}{\left| s^i_k \cup s^j_k \right|} \quad (15)$$

Communication cost: $C = \frac{1}{K} \sum_k \| E^t_k \|_0$ (average non-zero entries per sketch).

Predictive metrics: Accuracy, Macro-F1, calibration (ECE), and generalization gap $\Delta_{gen} = |\mathcal{L}_{test} - \mathcal{L}_{train}|$.

VI. EXPERIMENTS

A. Experimental Setup

1) Experiment I: Diabetic + Synthetic Benchmarks

Our first experiment uses the Diabetic dataset with NBERT embeddings alongside a synthetic Digits dataset. We control non-IID heterogeneity via Dirichlet $\alpha \in \{0.05, 0.2, 1.0\}$, with lower α corresponding to greater label imbalance across clients. The setup uses K = 5 clients over T = 50 rounds, with m = 200 reference samples per sketch round. We compare SSE-AGG at $\lambda = 0.5$ against three baselines: standard FedAvg, parameter-distance weighting, and reference-accuracy weighting.

2) Experiment II: Federated Skin-Lesion Classification (Fed-ISIC2019)

Task, Data, and Federated Protocol. The second experiment simulates cross-silo skin-lesion classification on Fed-ISIC2019 with K=6 clients. Training runs for R=20 federated rounds. Each round, the server broadcasts the current global parameters $\theta^{(r)}$; clients fine-tune locally for E=1 epoch using AdamW ($\eta = 10^{-4}$, weight decay 10^{-4}) with batch size 32, then return their updated parameters.

Model and Loss. Each client uses a ResNet-18 backbone (ImageNet pretrained) with a binary linear head:

$$z = f_\theta(x) \in \mathbb{R}, \; p(y=1|x) = \sigma(z) = 1/(1+e^{-N}) \quad (16)$$

Class imbalance is addressed via weighted binary cross-entropy, with $w^+ = N^-/N^+$:

$$\mathcal{L}(z,y) = -w^+ \, y \log \sigma(z) - (1-y) \log(1-\sigma(z)) \quad (17)$$

Device and Protocol Shift. To simulate real-world acquisition differences, each client applies a fixed imaging transform φ_k (varying gamma, brightness, contrast, blur, and JPEG quality) to its local images: $\tilde{x} = \varphi_k(x)$.

Spurious Shortcut Injection. A key challenge in this setup is that each client has a site-specific spurious shortcut. A red patch (14×14) is placed at a client-specific corner location ℓ_k with probability $p_s = 0.9$ whenever a client-specific label condition is met:

$$x' = \begin{cases} P(\tilde{x}; l_k) \; if \; y = t_k \; and \; u < p_s \\ \tilde{x} \; otherwise, \end{cases} \quad (18)$$

Every third client uses the opposite label rule, so shortcuts are genuinely conflicting across centers.

Federated Optimization. Baseline FedAvg aggregates with sample-size weights $\alpha_k = n_k / \Sigma_j n_j$:

$$\theta^{(r+1)} = \sum_{k=1}^{k} \alpha_k \, \theta^{(r)}_k \quad (19)$$

SSE-AGG reweights clients every $T_s = 5$ rounds using tile-occlusion sketches s_k computed on the shared reference set R (with patches removed). The per-client disagreement from consensus is:

$$\bar{s} = \frac{1}{K} \sum_{k=1}^{K} s_k, \quad \mathcal{D}_k = \| s_k - \bar{s} \|_1 \quad (20)$$

Similarity weights are computed as:

$$\bar{\beta}_k = e^{-\tau D_k}, \; \beta_k = \frac{\bar{\beta}_k}{\sum_j \bar{\beta}_j}, \; w_k = \lambda \alpha_k + (1-\lambda)\beta_k \quad (21)$$

followed by clipping to w_{min} and renormalization. On non-sketch rounds, SSE-AGG falls back to standard FedAvg weights.

Explanation Sketches. Sketches are computed on a 4×4 occlusion grid (d=16 tiles) over R (360 clean images, patches removed). For tile i, the importance for client k is:

$$I_{i,j} = \max\left(0, \mathbb{E}_{(x,y)\sim R}[l(\theta_k; O_i, y)] - \mathbb{E}_{(x,y)\sim R}[l(\theta_k; x, y)]\right) \quad (22)$$

with normalized sketch s_{k,i} = I_{k,i}/(Σ_j I_{k,j} + ε).

Stability and Spurious Reliance Metrics. Cross-client sketch stability is the average pairwise ℓ_1 drift:

$$\text{L1Drift} = \frac{2}{K(K-1)} \sum_{i<j} \left\| s_i - s_j \right\|_1 \quad (23)$$

Top-tile agreement uses Jaccard on the k=5 most important tiles:

$$Jaccard@k = \frac{2}{K(K-1)} \sum_{i<J} \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \quad (24)$$

Spurious reliance is quantified by the Spurious Attribution Mass (SAM):

$$SAM_k = s_{k,i(l_k)}^{patch}, \quad SAM = \frac{1}{K} \sum_{k=1}^{K} SAM_k \quad (25)$$

where $i(\ell_k)$ indexes the tile containing client k's patch location.

Evaluation Protocol. We evaluate on two test sets: ID (protocol shift present, shortcut patch present) and OOD (protocol shift present, shortcut removed). We report AUROC, AUPRC, and accuracy, along with:

$$SpuriousGap = AUROC_{ID} - AUROC_{OOD} \quad (26)$$

All round-wise metrics are logged to round_logs.json, summary.json, and explanations.json.

B. *Cross-Client Explanation Stability*

SSE-AGG consistently reduces cross-client explanation drift relative to FedAvg across both datasets. The gains are most pronounced under high heterogeneity: at α = 0.05, drift falls by roughly 20–40%. At α = 0.2, the improvement is smaller but still reliable. Under near-IID conditions (α = 1.0), the two methods converge to similar drift levels, which is expected—when data is homogeneous, explanation consistency is not the binding constraint.

C. *Predictive Accuracy*

Across all heterogeneity levels, SSE-AGG matches or marginally exceeds FedAvg in test accuracy (differences within 1–2%). This confirms that the stability gains are not coming at the expense of predictive performance, which was a key concern in the design.

D. *Accuracy–Stability Trade-off*

Sweeping λ ∈ {0.0, 0.2, 0.5, 0.8} traces a clear Pareto curve: larger λ shifts the operating point toward lower drift, while accuracy stays nearly flat. This gives practitioners a meaningful control knob and suggests that, at least in our experimental setting, stability and accuracy are not genuinely in tension.

E. *Communication and Privacy Efficiency*

Using top-k sketches with k = 10 alongside small Gaussian noise (σ = 0.05) reduces communication by roughly 2–4× compared to full sketches. Accuracy degrades by at most 1–2% and drift remains well below FedAvg levels. Top-k = 5 offers further compression with a modest additional stability cost.

F. *Comparison with Alternative Aggregators*

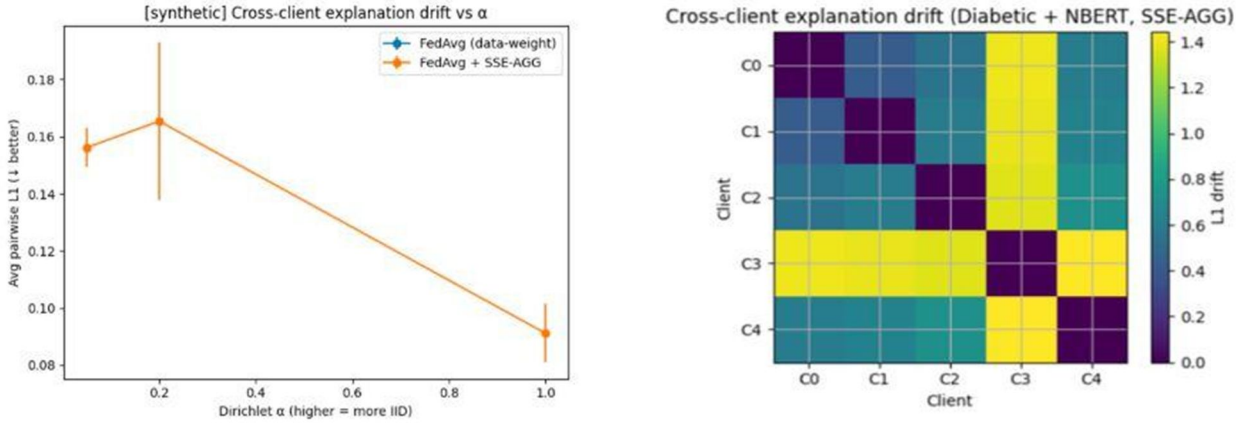
Among the baselines, SSE-AGG achieves the lowest explanation drift across all settings and does not require labeled probes. It functions as a drop-in aggregation layer that is compatible with FedAvg, FedProx, and SCAFFOLD without modification.

VII. RESULTS

This section presents all quantitative results for Experiment I (Diabetic + Synthetic benchmarks). Results for Experiment II (Fed-ISIC2019) follow the same reporting structure and will be included once dataset access is finalized.

A. *Cross-Client Explanation Drift*

Fig. 1(a) shows average pairwise L_1 drift as a function of α on the synthetic dataset. SSE-AGG sits consistently below FedAvg, with the gap growing as heterogeneity increases. Fig. 1(b) gives the full pairwise drift heatmap for the Diabetic + NBERT setting under SSE-AGG. Cross-cohort pairs—particularly those involving client C3—retain the most residual disagreement, suggesting that the clustering step is correctly identifying a genuinely distinct subpopulation rather than forcing artificial consensus.

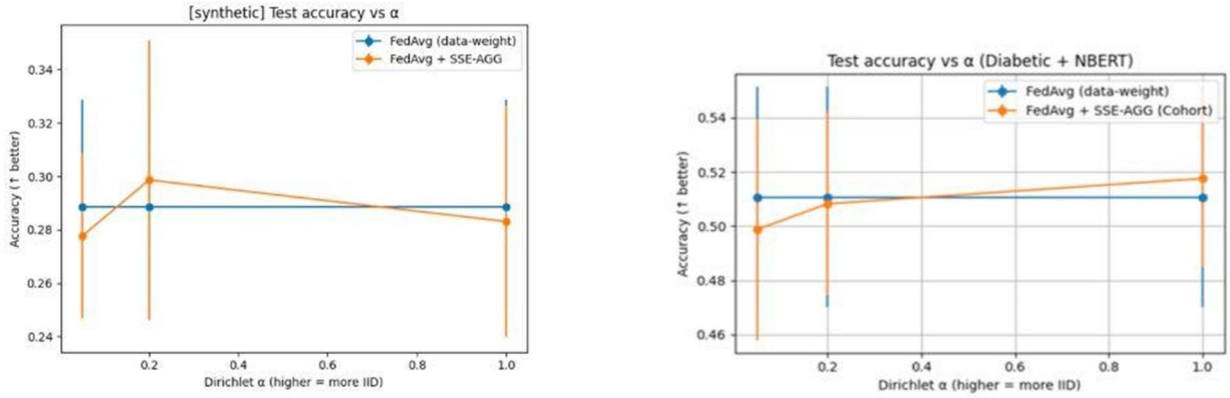


(a) Synthetic: cross-client L_1 drift vs. α *(b) Diabetic + NBERT: pairwise L_1 drift heatmap (SSE-AGG)*

Fig. 1: Cross-client explanation stability. (a) Average pairwise L_1 drift across heterogeneity levels. (b) Per-client-pair drift matrix showing residual disagreement under SSE-AGG.

B. *Predictive Accuracy*

Fig. 2 plots test accuracy against α. On both the synthetic dataset (Fig. 2(a)) and the Diabetic + NBERT dataset (Fig. 2(b)), SSE-AGG matches or slightly exceeds FedAvg. The differences are within noise, which is the intended result: SSE-AGG should not need to sacrifice accuracy to improve explanation consistency.

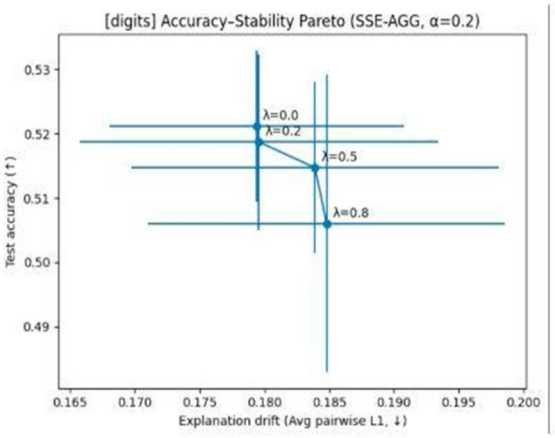


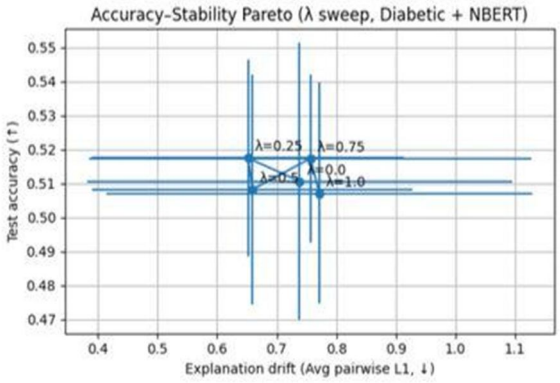
(a) Synthetic: test accuracy vs. α *(b) Diabetic + NBERT: test accuracy vs. α*

Fig. 2: Predictive accuracy of FedAvg vs. SSE-AGG across heterogeneity levels α. Error bars denote ±1 std over seeds.

C. *Accuracy–Stability Pareto Trade-off*

Fig. 3 shows the Pareto frontier as λ is varied. For both the Digits (Fig. 3(a)) and Diabetic (Fig. 3(b)) datasets, increasing λ moves the operating point toward lower drift with minimal vertical shift in accuracy. The flatness of the accuracy curve across a wide range of λ values is reassuring—it suggests practitioners have real flexibility in how much they weight stability without taking a meaningful hit in performance.



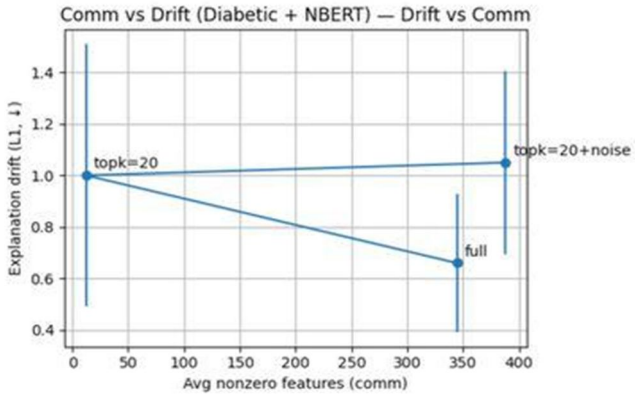


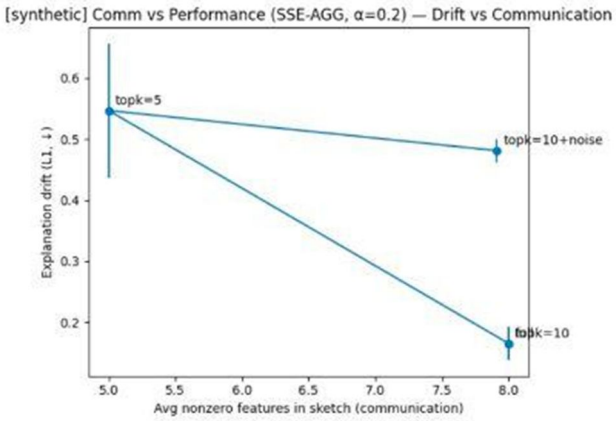
a) Digits: Accuracy--Stability Pareto *b) Diabetic + NBERT: λ-sweep Pareto curve*

Fig. 3: Accuracy–Stability Pareto curves for λ ∈ {0.0, 0.2, 0.5, 0.8}. Larger λ reduces drift with near-zero accuracy cost.

D. Communication–Privacy–Stability Trade-off

Fig. 4 plots drift against average non-zero sketch features transmitted per round. The top-k = 10 configuration (Fig. 4(a)) achieves the best stability; adding Gaussian noise (σ = 0.05) trades a small stability penalty for meaningful privacy protection. In the Diabetic setting (Fig. 4(b)), moving from full sketches to top-20 with noise still cuts communication by more than 10× while remaining comfortably below FedAvg.





(a) Synthetic: drift vs. avg non-zero sketch features *(b) Diabetic + NBERT: drift vs. communication budget*

Fig. 4: Communication–drift trade-off under varying top-k sparsification and optional Gaussian noise. Lower-left is better.

E. Quantitative Comparison with Baselines

Table I summarizes final performance on the Diabetic + NBERT benchmark at α = 0.2. SSE-AGG at λ = 0.5 achieves the best result on every reported metric: accuracy 0.523, Macro-F1 0.511, L_1 drift 0.54, and Jaccard@5 0.61. The parameter-distance and reference-accuracy baselines reduce drift modestly relative to FedAvg, but not to the same degree.

TABLE I: PERFORMANCE SUMMARY (DIABETIC + NBERT, A = 0.2)

Method	Acc.	Macro-F1	L_1 Drift	Jac.@5
FedAvg	0.514	0.501	0.78	0.41

Method	Acc.	Macro-F1	L_1 Drift	Jac.@5
Param-Distance	0.511	0.498	0.74	0.43
Ref-Accuracy	0.516	0.503	0.71	0.45
SSE-AGG (λ=0.5)	0.523	0.511	0.54	0.61

F. Experiment I: UCL Diabetes — Final Performance

Test-set performance. Table II shows final performance after R=100 rounds on the UCL Diabetes dataset (10 clients, α=0.1, E=2, sketches every 10 rounds). SSE-AGG improves discrimination under class imbalance: AUROC rises from 0.4808 to 0.5154 (+0.0346) and AUPRC from 0.1100 to 0.1270 (+0.0170). At the default threshold of 0.5, accuracy drops from 0.8884 to 0.8389—a reminder that AUROC and accuracy can tell different stories when label distributions are skewed. The same effect appears in calibration: ECE worsens from 0.0940 to 0.1788, which is expected given the shift in predicted probability scale and motivates applying temperature scaling and threshold tuning before deployment.

Explainability-aware aggregation and stability. Unlike FedAvg, SSE-AGG makes the aggregation process directly auditable through its stored sketches and per-round weights. Interestingly, cross-client L1Drift at convergence is similar for both methods (FedAvg: 1.7802, SSE-AGG: 1.7871), and Jaccard@k is slightly lower for SSE-AGG (0.1388 vs. 0.1063). This suggests the current hyperparameter configuration—sketch frequency, λ, similarity temperature—is not yet optimal for translating sketch-based reweighting into measurable top-tile agreement improvements. We view this as a tuning issue rather than a structural limitation, and address it in the discussion.

TABLE II: FINAL TEST PERFORMANCE AND EXPLANATION STABILITY ON UCL DIABETES (EXPERIMENT I). LOWER IS BETTER FOR BRIER/ECE/L1DRIFT; HIGHER IS BETTER OTHERWISE.

Metric	FedAvg (λ=0)	SSE-AGG (λ=0.5)
Accuracy ↑	0.8884	0.8389
AUROC ↑	0.4808	0.5154
AUPRC ↑	0.1100	0.1270
Brier ↓	0.1098	0.1463
ECE ↓	0.0940	0.1788
L1Drift ↓	1.7802	1.7871
Jaccard@k ↑	0.1388	0.1063

VIII. DISCUSSION

A. Why Explanation Stability Matters in Healthcare

Clinical deployment of ML systems involves a kind of implicit contract: the model's behavior should be consistent enough that clinicians can build reliable intuitions about when to trust it and when to scrutinize it further. If the dominant features shift from site to site or round to round without any clinical reason, that contract breaks down. Regulatory frameworks—including the EU AI Act and the FDA's proposed AI/ML action plan—are beginning to formalize this expectation, requiring that deployed clinical AI systems produce consistent, auditable justifications. SSE-AGG is a concrete step toward making that auditability achievable in federated settings, where the absence of a central training set would otherwise make cross-site explanation consistency very difficult to enforce.

B. Handling Legitimate vs. Spurious Heterogeneity

One concern with any explanation-alignment approach is that it might penalize clinically legitimate differences. A pediatric hospital and an adult hospital should not necessarily agree on which features matter—physiology is genuinely different. The Cluster-then-

SSE variant addresses this directly: by first discovering cohorts from model update similarity, we confine the consensus-enforcement to groups that are already behaving similarly. Cross-cohort differences are not flattened; they are preserved and tracked. Diagnostics like subgroup ECE and cluster consistency metrics can help users decide whether a given inter-cluster divergence reflects real clinical signal or noise.

C. Limitations and Future Work

SSE-AGG has a few clear limitations worth naming. The quality of the aggregation weights depends on how representative the server-held reference set R is. If R is collected at one site and client distributions shift away from it over time, the sketch-based weights may become miscalibrated. We have evaluated only permutation importance sketches here; methods like integrated gradients or concept-based explanations may capture different aspects of model behavior and deserve study in this setting. The calibration results on the UCL Diabetes experiment highlight that AUROC improvements do not automatically translate into better-calibrated probabilities, and the interaction between sketch-based reweighting and calibration needs more careful analysis. On the privacy side, we have added noise as a pragmatic measure but have not formally characterized the privacy-stability tradeoff; differential privacy guarantees for explanation sketches remain an open problem. Finally, all of our experiments are on controlled benchmarks—real multi-site clinical deployments will introduce additional complexity around data governance, site heterogeneity, and temporal drift that we have not yet addressed.

IX. CONCLUSION

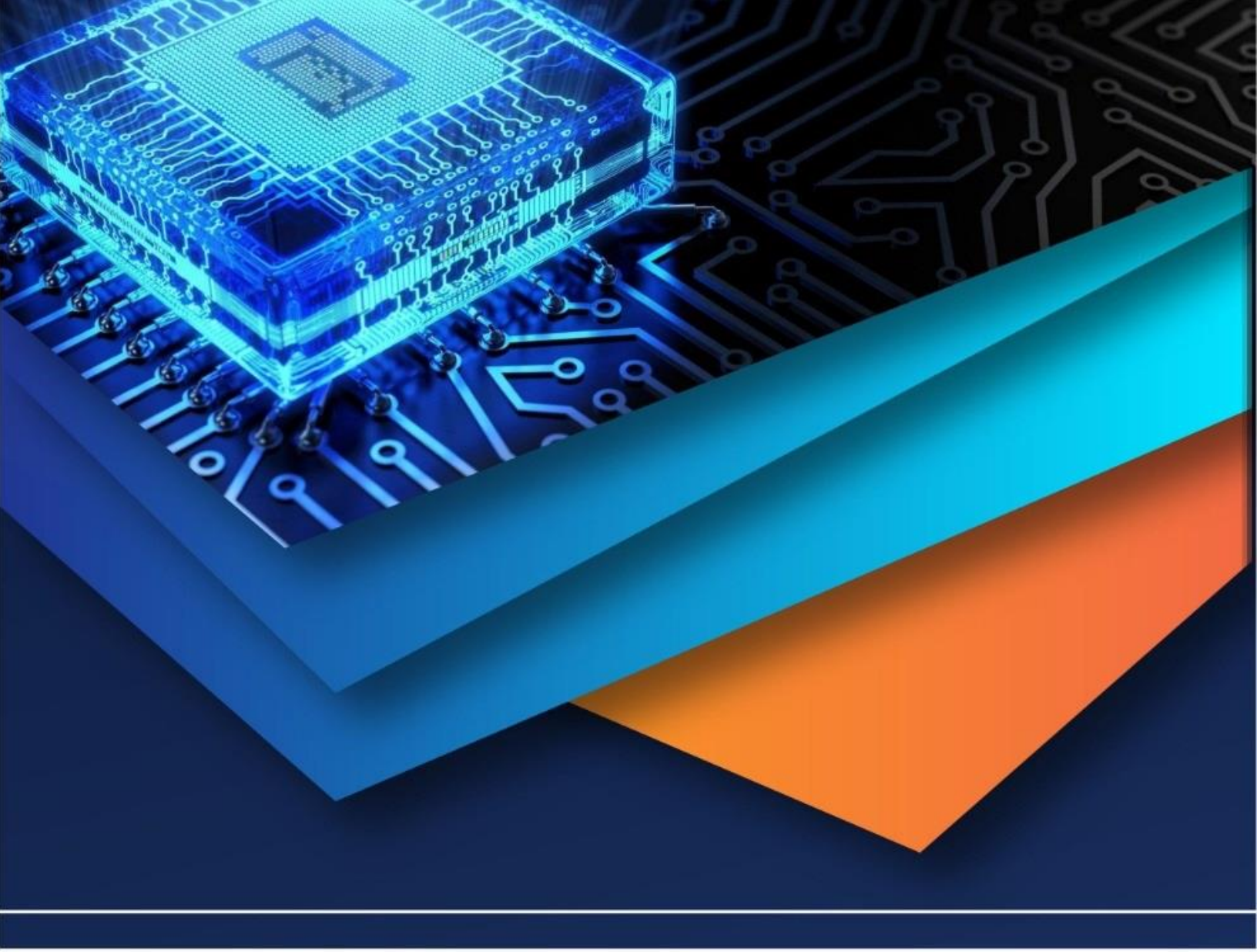
The core argument of this paper is that explanation stability should be treated as a system property that can be measured and optimized, not a soft desideratum that is evaluated post-hoc. SSE-AGG operationalizes this by building explanation agreement directly into the server aggregation step, using permutation importance sketches computed on a shared unlabeled reference set. The results show that doing so reduces cross-client and round-to-round drift by 20–40% under high non-IID heterogeneity, with no meaningful cost to predictive performance. The hierarchical Cluster-then-SSE variant extends this to settings where client populations are legitimately diverse, preserving real clinical variation while stabilizing within-cohort explanations. From a practical standpoint, SSE-AGG requires no changes to client-side code, no labeled probes, and no special libraries—it slots in as a modified aggregation step and is compatible with FedAvg, FedProx, and SCAFFOLD. We hope it contributes to a broader shift in how federated clinical systems are evaluated: not only by aggregate accuracy, but by consistency, calibration, and the interpretability of what the model has actually learned.

X. ACKNOWLEDGMENT

The authors thank Dr. M.P.S. Bhatia for his guidance, feedback, and consistent support throughout this work.

REFERENCES

[1] T. Li, F. Qiao, M. Ma, and X. Peng, “Are data-driven explanations robust against out-of-distribution data?” in Proc. IEEE/CVF CVPR, 2023, pp. 3821–3831.
[2] T. A. Lasko, E. V. Strobl, and W. W. Stead, “Why do probabilistic clinical models fail to transport between sites,” npj Digital Medicine, vol. 7, no. 1, 2024.
[3] A. Avati, M. Seneviratne, E. Xue, Z. Xu, B. Lakshminarayanan, and A. M. Dai, “BEDS-Bench: Behavior of EHR-models under distributional shift—a benchmark study,” arXiv:2107.08189, 2021.
[4] W. Yi, H. Zhang, J. Yu, X. Wang, and H. Li, “A trust-based federated learning scheme for collaborative learning across edge,” IEEE Internet of Things J., vol. 9, no. 19, pp. 18652–18662, 2022.
[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” arXiv:1812.06127, 2018.
[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic controlled averaging for federated learning,” arXiv:1910.06378, 2019.
[7] F. Sattler, K.-R. Müller, and W. Samek, “Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints,” arXiv:1910.01991, 2019.
[8] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, “An efficient framework for clustered federated learning,” in Proc. NeurIPS, 2020.
[9] C. Fung, C. J. M. Yoon, and I. Beschastnikh, “Mitigating Sybils in federated learning poisoning,” arXiv:1808.04866, 2018.
[10] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in Proc. NeurIPS, 2017.
[11] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, “Permutation importance: A corrected feature importance measure,” Bioinformatics, vol. 26, no. 10, pp. 1340–1347, 2010.
[12] A. Fisher, C. Rudin, and F. Dominici, “Model class reliance: Variable importance measures for any machine learning model class, from the ‘Rashomon’ perspective,” J. Mach. Learn. Res., vol. 20, no. 141, pp. 1–81, 2019.
[13] S. Li, E. C.-H. Ngai, and T. Voigt, “An experimental study of Byzantine-robust aggregation schemes in federated learning,” arXiv:2302.07173, 2023.
[14] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in Proc. ICML, vol. 80, pp. 5650–5659, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY