Análise de estratexias de aprendizaxe federada robustas a datos heteroxéneos

Roi Martínez Enríquez[a,*], **Roberto Iglesias Rodríguez**[a,1] **e Francisco Javier García Polo**[a,2]

[a]CiTIUS. Universidade de Santiago de Compostela

Resumo. A aprendizaxe federada (AF) permite adestrar modelos de forma descentralizada sen compartir datos sensibles, mais sofre importantes limitacións en contextos con datos heteroxéneos. Os algoritmos tradicionais tenden a fallar na xeneralización global ou na personalización local. Este traballo presenta FLProtector, un marco dual no que cada cliente aprende un incremento local sobre un modelo global compartido, e decide dinamicamente que modelo utilizar en inferencia mediante un autoencoder adestrado localmente para detectar entradas fóra de distribución. Ademais, FLProtector incorpora un mecanismo de agregación robusto baseado na consistencia de gradientes, que atenúa a influencia de clientes con actualizacións que se desvían do rumbo global agardado. A proposta é avaliada baixo distintos niveis de heteroxeneidade no benchmark DigitFive, mostrando melloras consistentes fronte métodos clásicos e de personalización do estado da arte, logrando un equilibrio efectivo entre personalización e xeneralización. O sistema mantén o seu rendemento incluso ante clientes estritamente maliciosos, e o estudo por capas confirma a relevancia de cada unha das súas compoñentes. Por último, a proposta salienta por non requerir unha sintonización sensible de hiperparámetros, o que facilita a súa aplicabilidade en escenarios reais.

1. Introdución

A aprendizaxe federada (AF) xorde como un paradigma orientado á preservación da privacidade, que permite a clientes descentralizados adestrar modelos de aprendizaxe automática de forma colaborativa sen necesidade de compartir datos en bruto (véxase figura 1). Con todo, en escenarios realistas, os clientes adoitan dispoñer de datos procedentes de distintas fontes – por exemplo, debido a variacións no comportamento dos usuarios, os dispositivos de adquisición ou os propios contornos –, o que dá lugar a datos non independentes nin identicamente distribuídos *non-IID* durante o proceso de federamento. Esta heteroxeneidade plantexa desafíos importantes para os algoritmos de AF estándar, como FedAvg [10], que asumen un obxectivo común e unha representación de datos uniforme. Como resultado, os modelos globais adestrados baixo estas hipóteses adoitan presentar un rendemento deficiente e unha limitada capacidade de personalización.

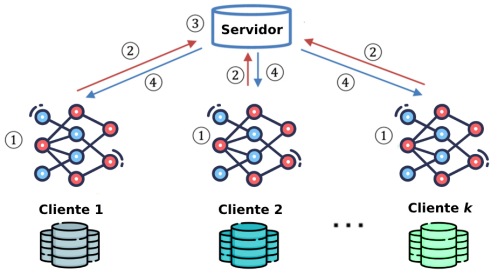

Figura 1. Esquema do proceso tradicional de aprendizaxe federada: en cada ronda, os clientes adestran de maneira local cadanseu modelo e envían as actualizacións ao servidor en forma de gradientes (sen necesidade de compartir datos), que realiza a agregación e devolve o modelo global actualizado.

Para abordar estes retos, a Aprendizaxe Federada Personalizada consolidouse como unha liña de investigación activa, que ten como obxectivo equilibrar dúas metas frecuentemente en conflito: manter un modelo global compartido que aproveite o coñecemento colectivo, e adaptar dito modelo á distribución específica de datos de cada cliente. Unha das estratexias máis salientables neste ámbito é a *interpolación de modelos*, onde cada cliente mantén tanto un modelo global como un personalizado (local). Esta estratexia permite combinar de forma flexible patróns globais con especialización nas particularidades locais.

A interpolación de modelos é particularmente poderosa porque relaxa a estrita suposición de que todos os clientes deben converxer a un único óptimo global. En contornos heteroxéneos, forzar a todos os clientes a acordar un modelo unificado a miúdo conduce a un rendemento subóptimo para a maioría. En cambio, a inter-

* Autor principal. Email: roi.martinez@usc.es
[1] Titor
[2] Cotitor

polación proporciona un espectro continuo entre a xeneralización completa e a personalización completa, permitindo a cada cliente atopar un punto de equilibrio óptimo en función dos seus datos. Métodos como APFL [3] adoptaron este principio, logrando melloras significativas sobre os enfoques clásicos de AF.

Con todo, a pesar do seu atractivo teórico, unha limitación crítica permanece sen resolver: *cando* e *como* debe un cliente decidir que modelo usar no momento da inferencia? A maioría dos métodos actuais combinan modelos de forma estática ou aplican un coeficiente de interpolación fixo durante o adestramento e a implementación. Estes enfoques asumen que as características de distribución dos datos de entrada do test son coñecidas e similares aos datos de adestramento, unha suposición forte e a miúdo irrealista en escenarios prácticos.

Neste traballo proponse un novo marco de AF de dous modelos que inclúe un mecanismo explícito para decidir, para cada entrada, se debe aplicar o modelo global ou o personalizado. Esta decisión tómase de forma dinámica no momento da inferencia utilizando un detector de novidade específico para cada cliente baseado en autoencoders. O modelo personalizado adóptase como un incremento sobre o global, nun bucle de coevolución, garantindo unha aliñación continua. Cando se detecta que unha entrada do test está "dentro da distribución", utilízase o modelo personalizado; doutro modo, o modelo global ten prioridade, ofrecendo unha mellor xeneralización.

Ademais, introdúcese un mecanismo de ponderación de gradientes baseado na consistencia para refinar a agregación global. Usando un enfoque inspirado en L-BFGS [1], detéctase e redúcese a influencia dos clientes cuxas actualizacións se desvían consistentemente da traxectoria de optimización esperada, o que normalmente indica un comportamento malicioso ou unha heteroxeneidade extrema dos datos.

A proposta é avaliada no desafiante *benchmark Digit-Five*, tanto en configuracións de clientes totalmente como parcialmente heteroxéneas. O método xeralmente supera as referencias estándar e métodos de AFP de última xeración en rendemento local, así como a nivel global ante casos de forte heteroxeneidade de datos. Ademais, non require unha sutil configuración de hiperparámetros, o que mellora a súa robustez e facilita a súa implementación en aplicacións do mundo real.

As contribucións deste traballo pódense resumir do seguinte xeito:

- Proposta dun marco de AF de dous modelos cunha selección de modelos dinámica a través da detección de novedad específica para cada cliente, permitindo unha inferencia adaptativa no momento da proba.
- Deseño dun esquema de adestramento de coevolución no que se aprenden incrementos personalizados sobre un modelo global compartido para manter a aliñación global e a personalización local.
- Introdución dun mecanismo de ponderación de gradientes baseado na consistencia das actualizacións para mellorar a robustez contra clientes atípicos.
- Demostración dun rendemento robusto nun *benchmark* multi-dominio, superando os métodos de AFP existentes en configuracións con distintos graos de heteroxeneidade.

2. Estudo do estado da arte

A aprendizaxe federada espertou un grande interese debido á súa capacidade para adestrar modelos de aprendizaxe automática de forma colaborativa utilizando conxuntos de datos descentralizados, preservando ao mesmo tempo a privacidade dos datos dos usuarios. Así a todo, o algoritmo estándar de AF, **FedAvg** [10], sofre unha degradación significativa do rendemento cando os conxuntos de datos dos clientes presentan unha alta heteroxeneidade. Unha revisión exhaustiva dos enfoques existentes de aprendizaxe federada personalizada que xurdiron para abordar estas limitacións pódese atopar en [15]. En xeral, os enfoques existentes poden agruparse en dúas liñas principais de investigación: métodos *baseados en regularización* e os métodos de *interpolación de modelos*. Os primeiros abordan a heteroxeneidade introducindo termos de penalización nas funcións de perda locais, restrinxindo así as actualizacións dos clientes para que permanezan próximas ao modelo global. Por outra banda, os métodos de interpolación manteñen dous modelos por cliente (global e local), permitindo un compromiso entre xeneralización global e especialización local mediante a interpolación de parámetros.

FedProx [8] introduce unha regularización proximal na optimización local, penalizando as desviacións significativas respecto ao modelo global durante o adestramento. Se ben esta estratexia mellora a estabilidade da aprendizaxe en escenarios *non-IID*, o seu rendemento depende fortemente do valor do hiperparámetro de regularización μ. Estudos empíricos mostran que o rendemento é altamente sensible a μ, requirindo unha sutil configuración para acadar resultados satisfactorios.

SCAFFOLD [6] é outro método baseado en regularización que se centra na redución da variabilidade nas actualizacións dos clientes mediante o uso de variables de control que reducen a varianza nas actualizacións locais. Estas variables corrixen o desvío dos clientes, aliñando os gradientes locais co obxectivo global e, así, mellorando a converxencia en distribucións heteroxéneas.

Por outra banda, a interpolación explícita de modelos gaña terreo con métodos como **APFL** (Adaptive Personalized Federated Learning) [3], que adestra de forma concurrente un modelo global e un local por cliente, combinándoos mediante un coeficiente de interpolación axustable. Aínda que APFL ofrece maior flexibilidade que FedProx ou SCAFFOLD, presenta limitacións notables: o coeficiente de interpolación tende a ser estático ou a adaptarse moi lentamente e, ademais, carece de estratexia explícita para decidir que modelo aplicar segundo a distribución de entrada en tempo de inferencia. Isto pode afectar negativamente ao rendemento en escenarios con alta variabilidade ou incerteza.

Mais alá da interpolación, outro problema clave no contexto da AF é a detección de clientes anómalos ou altamente diverxentes, especialmente en contextos adver-

sarios ou con forte heteroxeneidade. Nesta liña, **FLDetector** [14] propón unha aproximación baseada en gradientes de segunda orde mediante o algoritmo *L-BFGS* (Limited-memory BFGS), que estima as actualizacións esperadas e detecta desviacións significativas para identificar clientes maliciosos ou atípicos. Con todo, FLDetector actúa principalmente como ferramenta de diagnóstico —é eficaz na detección, pero non proporciona recomendacións directas sobre como ponderar ou adaptar as actualizacións dos clientes de forma dinámica.

Fronte a estas limitacións, a proposta deste traballo sintetiza as ideas anteriores e extende as súas capacidades. Partindo dun marco de dobre modelo, introdúcese un mecanismo explícito e dinámico: un sistema de detección de novidade por cliente mediante autoencoders, que permite seleccionar, para cada mostra, entre o modelo global e o personalizado en tempo de inferencia. Esta selección baseada na novidade permite unha personalización máis robusta e precisa, superando a interpolación estática ou heurística que limita os métodos existentes. Ademais, este método é notablemente menos dependente da sutil configuración de hiperparámetros en comparación con estratexias baseadas en regularización como FedProx, mellorando a aplicabilidade práctica.

Adicionalmente, integrando as ideas de FLDetector, introdúcese un mecanismo baseado na consistencia de gradientes utilizando aproximacións de L-BFGS, non só para a detección de anomalías de clientes senón para variar activamente os pesos de agregación das actualizacións de cada cliente e reducir sistematicamente a influencia de clientes consistentemente diverxentes durante a agregación global.

Así, a aproximación proposta combina de forma única a interpolación dinámica de modelos explícita cunha estratexia de ponderación de clientes adaptativa, proporcionando unha solución máis holística ao desafío da personalización en condicións de alta heteroxeneidade que os métodos anteriores abordan só de maneira parcial ou indirecta.

3. A proposta: FLProtector

3.1. Modelo global federado + Incremento local

Como se introduciu anteriormente, este traballo propón unha estratexia de aprendizaxe federada personalizada que se basea nun proceso cíclico composto por dúas operacións fundamentais en cada ronda: unha fase de adestramento federado (mediante o algoritmo clásico de agregación, cun número reducido de T_l épocas locais) e unha fase de aprendizaxe do modelo personalizado (personalización local). Como se observa no algoritmo 1, cada cliente pode reter dous modelos: o global –federado– representado polos pesos w, e o incremento local, representado polos pesos δ, de modo que: $personalized\_model(v) = federated\_model(w) + local\_increment(\delta)$

Os pesos do modelo global w^0 son inicializados de forma aleatoria e a personalización δ_i^0 de cada cliente establécese en cero. Durante o proceso de adestramento, altérnanse dúas fases principais en cada iteración do ciclo:

Algoritmo 1: Entrenamento Federado e Personalizado

entrada : N clientes, cada un con M_i mostras. Número de épocas locais T_l, número de épocas para entrenar o incremento T_i, número total de rondas T.
saída : Para cada cliente, modelo global W^T e modelos personalizados v_i^T
notación: f_i representa a función de perda local, ξ_i^t un minibatch local de datos, w_i é a copia local dos pesos globais

1 Inicializar w^0 aleatoriamente
2 Inicializar δ_i^0 a cero para todo cliente i
3 **for** t *in* $\{0,\dots,T\}$ **do**
4 **for** *cada cliente i* **do**
5 Copia do modelo global:
6 $w_i^{t,0} = W^t$
7 Adestramento local do modelo global:
8 **for** t_l *in* $\{0,\dots,T_l\}$ **do**
9 $w_i^{t,t_l+1} = w_i^{t,t_l} - \eta_t \nabla f_i(w_i^{t,t_l}; \xi_i^t)$
10 Promediado global no servidor: $W^{t+1} \leftarrow \sum_{i=1}^{N} \frac{M_i}{M} w_i^{t,T_l}$
11 Adestramento local do incremento: modelo personalizado
12 **for** *cada cliente i* **do**
13 Construír o modelo personalizado:
14 $v_i^{t+1} = W^{t+1} + \delta_i^t$
15 Adestramento local do modelo personalizado:
16 **for** t_i *in* $\{0,\dots,T_i\}$ **do**
17 $v_i^{t,t_i+1} = v_i^{t,t_i} - \eta_t \nabla f_i(v_i^{t,t_i}; \xi_i^t)$
18 Gardar o incremento para a seguinte iteración do ciclo:
19 $\delta_i^{t+1} = v_i^{t+1,T_i} - W^{t+1}$
20 **return** Modelo global: W^T, modelos personalizados: $v_i^T = W^T + \delta_i^T$

- **Adestramento federado**: Primeiro, o modelo global actualízase localmente durante T_l épocas (liñas 4-9). A continuación, cada cliente envía o seu modelo ao servidor, que realiza a agregación (liña 10).
- **Personalización local**: Cada cliente carga a súa rede local personalizada v_i^{t+1} como a suma do modelo global w^{t+1} máis o último incremento computado: δ_i^t (liña 14). A rede personalizada adéstrase con datos locais durante T_i épocas (liñas 15-17) e, finalmente, actualízase o valor do componente de personalización, incremento δ_i^{t+1} (liña 19).

É salientable neste esquema o establecemento dunha co-evolución dos modelos personalizados xunto co global, que permite a cada cliente **especializarse de maneira progresiva sen perder aliñación co coñecemen-**

to global compartido.

3.2. *Selección dinámica de modelos baseado en autoencoders*

Segundo a proposta presentada, cada cliente mantén dous modelos: un global e un personalizado. Isto plantexa de forma natural a cuestión de cando aplicar cada un deles durante a inferencia. Para resolver esta decisión adóptase un enfoque de detección de novidade baseado en autoencoders [13]. Este mecanismo permite que cada cliente avalíe se unha mostra de entrada é similar aos datos previamente vistos –caso no que se emprega o modelo personalizado– ou se se trata dunha mostra fóra de distribución, no que se recorre ao modelo federado global.

Tamén se exploraron estratexias alternativas, incluíndo métodos baseados en características como PCA ou representacións obtidas mediante redes CNN, combinadas con algoritmos de detección de novidade (por exemplo, One-Class SVM ou Local Outlier Factor) e enfoques de incerteza do modelo como o dropout bayesiano. Porén, estas alternativas obtiveron un rendemento inferior en comparación co método baseado en autoencoders.

Os autoencoders son redes neuronais adestradas para reconstruír a súa entrada, o que lles permite aprender unha representación comprimida da distribución subxacente dos datos. Nesta proposta, cada cliente adestra de forma independente un autoencoder utilizando o seu conxunto de datos local e calcula o erro medio de reconstrucción sobre as mostras de adestramento. Durante a inferencia, as novas entradas son pasadas a través do autoencoder, e **se o erro de reconstrucción supera a media de adestramento máis tres desviacións estándar, a entrada clasifícase como fóra de distribución**. Nestes casos, utilízase o modelo global para a predición en lugar do personalizado.

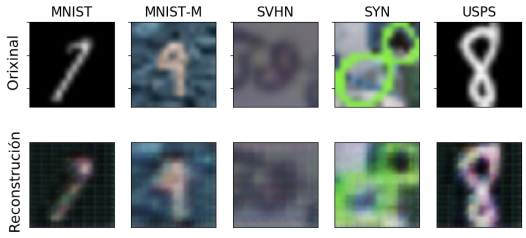

Figura 2. Exemplos de reconstrución xerados polos autoencoder para cada un dos cinco conxuntos de Digit-Five.

3.3. *Combinación lineal influenciada pola consistencia nas actualizacións do modelo*

No algoritmo estándar de *FedAvg* a contribución de cada cliente ao modelo global baséase unicamente na porcentaxe normalizada de datos en cada cliente (liña 8). Non obstante, pode darse o caso de que unha minoría de clientes posúa datos locais moi *non-IID* ou significativamente distintos, mentres que a maioría comparte características comúns na tarefa que se está aprendendo, ou presenta conxuntos de datos locais altamente correlacionados. Neste contexto, aplicar un peso uniforme baseado só na cantidade de datos pode levar a agregacións pouco representativas ou mesmo prexudiciais.

Para paliar esta situación, introdúcese un segundo sistema de ponderación que **reduce a influencia dos clientes discrepantes e reforza a contribución daqueles que mostran un comportamento máis coherente coa dinámica global**. Deste modo, os clientes maioritarios tenden a ter incrementos locais pequenos, mentres que os clientes máis disonantes manteñen un incremento maior que reflicte a súa diverxencia.

Inspirándose nos traballos de Zhang et al. sobre detección de clientes maliciosos (FLDetector [14], e FedRecover [2]), adóptase unha técnica para avaliar a consistencia das actualizacións de cada cliente con respecto ao comportamento global. Denotando como w_t os pesos correspondentes ao modelo global actual na iteración t, e como g_i ao gradiente calculado polo cliente i: $g_i = \nabla f(D_i, w)$, é posible estimar a súa próxima actualización (gradiente) Hessiana integrada para ese cliente específico:

$$g_i^t = g_i^{t-1} + H_i^T (w_t - w_{t-1})$$

Para simplificar, emprégase $\hat{H}^t$, unha aproximación compartida para todos os clientes da Hessiana global no instante t (véxase a ecuación 1), que se integra vía L-BFGS. Este algoritmo permite aproximar o produto vectorial Hessiano $H\Delta w$ empregando un historial limitado de gradientes e actualizacións pasadas, evitando así o custo computacional de calcular a Hessiana completa.

$$\hat{g}_i^t \approx g_i^{t-1} + \hat{H}^T (w_t - w_{t-1}) \tag{1}$$

Deste modo, obtense para cada cliente a estimación do seu gradiente esperado, $\hat{g}_i^t$, que se compara co gradiente real g_i^t calculado durante a fase de adestramento local, e mídese a súa discrepancia mediante a distancia euclidiana:

$$d^t = [\|\hat{g}_1^t - g_1^t\|_2, \|\hat{g}_2^t - g_2^t\|_2, ..., \|\hat{g}_n^t - g_n^t\|_2,]$$

Estas discrepancias normalízanse ($\hat{d}^t = d^t / |d^t|_1$) e acumúlanse ao longo de N iteracións para cada cliente:

$$s_i^t = \frac{1}{N} \sum_{r=0}^{N-1} \hat{d}_i^{t-r} \tag{2}$$

Unha alta puntuación de discrepancia indica que a actualización do cliente diverxe sistematicamente da traxectoria esperada, o que pode deberse a comportamento malicioso ou a unha distribución local moi distinta. Para reflectir esta confianza na fase de agregación defínense os pesos β_i^t mediante unha transformación tipo *softmax* invertida:

Sexa $\mathbf{d}^t = [d_1^t, d_2^t, \ldots, d_N^t]$ o vector de discrepancias para todos os N clientes na ronda t, o peso asignado a cada cliente calcúlase como:

$$\beta_i^t = \frac{\exp(-d_i^t)}{\sum_{j=1}^{N} \exp(-d_j^t)} \tag{3}$$

Esta transformación, respecto posibles alternativas lineares, **evita comprimir excesivamente o rango de valores** e permite unha diferenciación máis sensible entre clientes. Os pesos β_i^t resultantes son integrados no proceso de agregación global como:

$$w^{t+1} = \sum_{i=1}^{N} \left(X \cdot \frac{M_i}{M} + (1 - X) \cdot \beta_i^t \right) w_i^t \quad \text{con } X \in [0,1] \tag{4}$$

Aquí, M_i é o número de mostras do cliente i, $M = \sum_{i=1}^{N} M_i$, e X é un hiperparámetro escalar que controla a compensación entre o peso por tamaño do conxunto de datos e o peso por consistencia da actualización do modelo. Cando $X = 1$, a regra redúcese a *FedAvg* estándar; cando $X = 0$, só as puntuacións de consistencia determinan os pesos de agregación.

4. Análise experimental

O marco FLProtector proposto é avaliado baixo diferentes niveis de heteroxeneidade de datos, co obxectivo de analizar a súa capacidade para equilibrar a personalización e xeneralización. Os experimentos desenvólvense sobre o *benchmark Digit-Five*, que inclúe cinco conxuntos de datos de clasificación de díxitos con diferentes características visuais.

FLProtector compárase fronte métodos clásicos de AF, como *FedAvg*, *SCAFFOLD* e *FedProx*, e tamén fronte a *APFL*, técnica do estado da arte da AF personalizada. Ademais, realízase un estudo por capas para avaliar de forma illada a contribución de cada compoñente do marco proposto. Esta sección estrutúrase como segue: en primeiro lugar, descríbese o conxunto de datos e a configuración dos experimentos; a continuación, preséntanse os resultados do estudo por capas; e por último, detállase a comparación cos outros métodos.

4.1. Colección Digit-Five

Digit-Five é unha compilación dos datos de cinco conxuntos de datos de díxitos: MNIST [7], MNIST-M [4], SVHN [11], Synthetic Digits (SYN) [12], e USPS [5]. Cada conxunto de datos contén díxitos do 0 ao 9, pero presentan diferentes características e estilos, como se mostra na Figura 3. Por exemplo, os conxuntos de datos MNIST e USPS son en branco e negro, mentres que os outros son en cor. Esta colección de varios conxuntos permite a simulación dunha tarefa *non-IID*, onde o obxectivo segue sendo o mesmo (clasificar díxitos), pero os datos que cada cliente manexa poden variar, complicando o proceso de aprendizaxe colaborativa.

Dado que se necesita a mesma topoloxía de rede para realizar a aprendizaxe federada, e cada conxunto de datos ten diferentes dimensións de imaxe, todas as imaxes son transformadas ao mesmo tamaño: 16x16x3 [3].

[3] Para imaxes en branco e negro, a capa orixinal é triplicada para simular unha imaxe en cor (RGB) sen alterar o contido orixinal

Figura 3. Exemplos das imaxes dos diferentes conxuntos de datos de Digit-Five

4.2. Configuración dos experimentos

En todos os experimentos participan cinco clientes ($N = 5$), cada un dos cales ten acceso só aos seus propios datos. Cada cliente conta con 7000 mostras (imaxes) para adestramento e outras 2000 mostras do mesmo conxunto de datos para test. Todos os clientes implementaron unha rede neuronal convolucional (CNN) composta por catro capas convolucionais seguidas de catro capas totalmente conectadas.

Os experimentos realízanse sobre un número fixo de rondas de comunicación, con cada ronda consistindo en 5 épocas de adestramento local por cliente. Por defecto, todos os métodos son adestrados durante $T = 20$ rondas de comunicación global utilizando o optimizador Adam cunha taxa de aprendizaxe de 1×10^{-3}. Faise unha excepción para SCAFFOLD, cuxo mecanismo de redución de varianza depende de termos de corrección que están teóricamente fundamentados no uso de Stochastic Gradient Descent (SGD). Neste caso, emprégase SGD cunha taxa de aprendizaxe de 1×10^{-2}, e o adestramento é ampliado a 25 rondas de comunicación global para compensar a converxencia máis lenta que se observa normalmente con SGD en comparación con Adam.

Para garantir a equidade na comparación, todos os métodos son adestrados co mesmo número de actualizacións locais por ronda, e comparten unha inicialización e particionamento de datos consistentes a través das varias execucións.

4.3. Contornos de avaliación

Os rendementos son avaliados sobre dous conxuntos de test:

- **Local**: Formado polas 2000 mostras asignadas a cada cliente individual. Isto mide o rendemento personalizado en datos en distribución.
- **Global**: Resultado da unión dos conxuntos de test local de todos os clientes. Isto simula un escenario de implementación onde as entradas poden orixinarse de calquera dominio, incluídos aqueles non vistos localmente.

Para validar os experimentos, considéranse tres escenarios:

- `5b` - **Heteroxeneidade completa**: Cada un dos cinco clientes ten un conxunto de datos diferente, o que fai que a colaboración sexa subóptima.
- `d5` - **Heteroxeneidade parcial**: Catro clientes comparten o mesmo conxunto de datos, mentres que o quinto ten un diferente. Este escenario espera favorecer aos clientes maioritarios, mentres que o cliente minoritario pode beneficiarse menos da colaboración.
- `Ataques`: Este escenario explora a robustez do sistema ante posibles ataques, onde se simulan comportamentos maliciosos por parte dos clientes.

  Backdoor: O cliente malicioso duplica e modifica parte das imaxes de adestramento insertando un patrón visual distintivo, que é asociado deliberadamente cunha clase obxectivo, de modo que o modelo aprende a clasificar o patrón de forma errónea pero sistemática.

  Label Flip: Baséase na reasignación aleatoria de etiquetas nos datos de adestramento, para introducir ruído e confusión que deterioren o rendemento xeral do modelo global.

  Mean: O cliente malicioso altera a actualización enviada ao servidor invertindo o signo dos gradientes (e escalando para lograr un maior efecto) o que induce ao modelo global a alonxarse da dirección óptima de aprendizaxe.

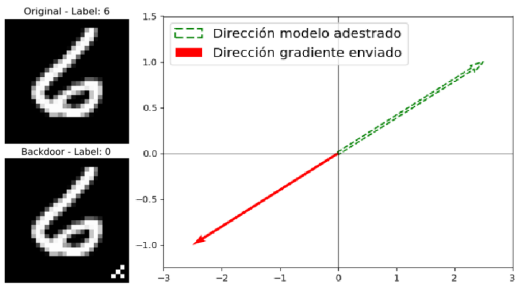

Figura 4. Ilustración dos ataques **Backdoor** (Á esquerda superior, a mostra orixinal de adestramento; debaixo, a mostra co patrón introducido e etiqueta cambiada ao obxectivo (neste caso 0)) e **Mean** (A gráfica da dereita mostra un esquema da dirección ideal do modelo e o gradiente oposto enviado polo cliente malicioso.)

É importante salientar que non todos os conxuntos de datos presentan o mesmo nivel de dificultade. Ao adestrar localmente, os conxuntos de datos como MNIST e USPS adoitan ofrecer un mellor rendemento en comparación cos outros. Para ilustrar os diferentes niveis de dificultade entre os conxuntos de datos, a táboa 1 mostra o rendemento obtido ao adestrar un modelo localmente en cada conxunto de datos, sen colaboración.

A partir destes resultados individuais, son deseñados os experimentos en contornos de heteroxeneidade. Dado que hai cinco conxuntos de datos dispoñibles, non é factible incluír todas as combinacións posibles de 4+1

MNIST	MNISTM	SVHN	SYN	USPS
96.9 %	81.1 %	73.3 %	85.9 %	98.6 %

Táboa 1. Resultados do adestramento individual en cada conxunto de datos utilizando 7000 imaxes de adestramento e 2000 imaxes de test

clientes, polo que se presenta unha configuración representativa dos posibles escenarios. En concreto, considéranse dous casos principais:

- **Maioría sinxela - MNIST** : Catro clientes utilizan un conxunto de datos máis fácil (MNIST), e o quinto cliente utiliza cada un dos restantes conxuntos de datos por separado.
- **Maioría difícil - SVHN** : Catro clientes utilizan un conxunto de datos máis difícil (SVHN), e o quinto cliente utiliza cada un dos restantes conxuntos de datos por separado.

Estas configuracións permiten investigar como a dificultade relativa do conxunto de datos maioritario influéncia no rendemento colaborativo global dun xeito controlado e interpretable.

4.4. Puntuacións de confianza

4.4.1. Sen ataques: heteroxeneidade natural

A figura 5 mostra a distribución das puntuacións de confianza asignadas a cada grupo de clientes (discrepante ou maioría) segundo a colección de orixe discrepante, en contornos `d5` con maioría SVHN. En todas as execucións, excepto aquela na que todos os clientes comparten SVHN, obsérvase un patrón claro: o cliente discrepante recibe puntuacións significativamente inferiores ao resto, o que reflicte a capacidade do sistema para detectar e cuantificar desviacións entre as propostas locais e a dinámica global. Ademais, as puntuacións practicamente idénticas no caso homoxéneo (todos SVHN) valida a neutralidade do mecanismo de ponderación cando non existe unha diverxencia real.

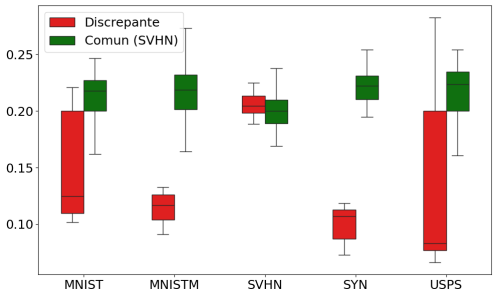

Figura 5. Puntuacións de confianza sen ataques segundo a colección do cliente discrepante (SVHN común).

A figura ilustra dúas formas distintas de diverxencia:

- No caso de **SYN (ou MNIST-M)** as puntuacións do discrepante son consistentemente baixas e con escasa dispersión, indicando que o sistema detecta de forma clara a súa incompatibilidade estrutural respecto ao consenso global.

- No caso de **MNIST, e especialmente USPS**, as puntuacións demostran unha maior variabilidade, o que suxire que o seu comportamento non é tan claramente discrepante, senón que fluctúa durante o adestramento.

Este fenómeno analízase en maior profundidade na figura 6. No caso de SYN obsérvase como a partir da segunda época os valores obtidos caen de forma abrupta ata o final do adestramento, mentres que USPS mostra un patrón máis gradual: comeza aliñado co grupo e mesmo con puntuacións superiores, pero a partir da sexta época a súa confianza descende ata estabilizar en valores baixos. Esta deriva progresiva indica que USPS, ao ser unha colección máis sinxela, converxe rapidamente e comeza a xerar actualizacións que difiren en magnitude ou dirección do resto. A análise dinámica das puntuacións de confianza revela que o sistema axusta progresivamente o seu nivel de confianza segundo o comportamento de cada cliente. Esto valida que as puntuacións de confianza non só son útiles para detectar ataques, senón tamén para xestionar desviacións naturais na aprendizaxe federada *non-IID*.

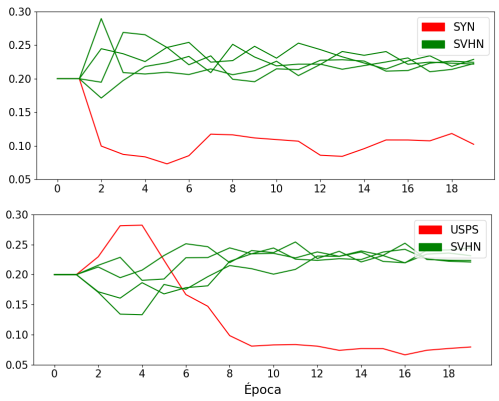

Figura 6. Evolución das puntuacións de confianza para os clientes discrepantes SYN (arriba) e USPS (abaixo) ao longo do adestramento con maioría SVHN.

4.4.2. *Con ataques*

A figura 7 presenta a distribución das puntuacións de confianza asignadas aos clientes participantes segundo o tipo de ataque perpetrado polo malicioso. O sistema, de novo, volve mostrar unha capacidade robusta de discriminación ante comportamentos anómalos.

Nos ataques *backdoor* e *mean* as puntuacións do atacante son practicamente nulas, o que implica a súa virtual exclusión do proceso de agregación. No caso de *label flip*, de natureza menos forte, o atacante obtén puntuacións máis moderadas pero consistentemente inferiores ao resto. Este comportamento corrobórase de maneira cuantitativa na táboa 2, que recolle as precisións obtidas polas diferentes técnicas de AF ante cada ataque. Pódese concluír que FLProtector preserva a precisión en clientes benignos en contornos limpos ao mesmo nivel que os métodos clásicos, mentres que é capaz de reducir

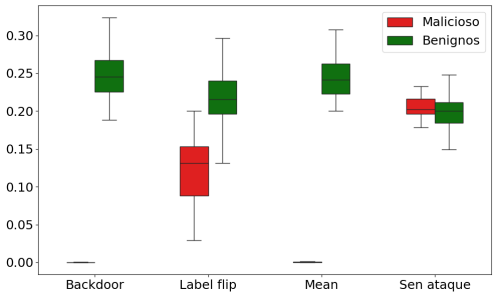

Figura 7. Puntuacións de confianza segundo o ataque do cliente malicioso (SVHN común).

o éxito de todos os ataques propostos de maneira consistente.

SVHN	None	Label flip	Mean	Backdoor
FedAvg	84.0 %	69.1 %	10.5 %	80.3 % (98 %)
Scaffold	81.2 %	79.3 %	10.8 %	80.2 % (98 %)
FedProx	**84.5 %**	75.0 %	7.7 %	83.2 % (98 %)
APFL	77.3 %	76.6 %	76.6 %	76.2 % **(3 %)**
FLProt.	80.7 %	**80.3 %**	**78.4 %**	79.1 % (4 %)

Táboa 2. Media das precisións obtidas polos clientes benignos baixo distintos tipos de ataque a SVHN.

4.5. *Análise da precisión dos encoders*

Unha das compoñentes chave da arquitectura proposta é o selector baseado en autoencoders, encargado de decidir para cada mostra se debe ser clasificada mediante o modelo global ou o personalizado. Para avaliar a súa eficacia, analizouse comportamento do autoencoder do cliente discrepante nun contorno `d5` con maioría SVHN.

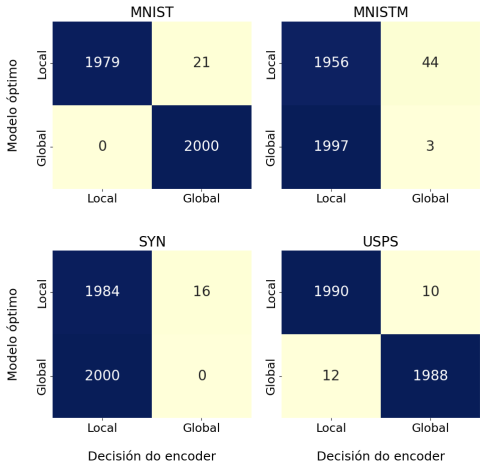

Figura 8. Decisións do encoder do cliente discrepante en contorno `d5` con maioría SVHN sobre test global (4000 mostras: 2000 SVHN e 2000 da colección do discrepante)

Foron construídas matrices de confusión para cada configuración, comparando a decisión tomada polo se-

lector coa elección óptima a posteriori, é dicir, o modelo (global ou local) adestrado para clasificar correctamente cada mostra no conxunto de test. Os resultados da figura 8 mostran que o selector:

- **Acerta de maneira sistemática na asignación de mostras ao modelo local** en todos os casos.
- **Comete erros na asignación ao modelo global** nas coleccións difíciles MNIST-M e SYN. Nestes casos, o selector tende a aplicar indebidamente o modelo local a mostras fóra da distribución de datos coñecida.

Estas observacións compleméntanse cos datos recollidos na táboa 3, que recolle os umbrais τ aprendidos por cada autoencoder. Apréciase unha clara correlación entre a complexidade do dominio e o valor de τ: en coleccións máis sinxelas como MNIST ou USPS, os umbrais son significativamente máis baixos que en coleccións máis complexas como MNIST-M ou SYN, onde a tarefa de trazar unha fronteira confiable entre o que debe ser tratado como coñecido ou non se volve máis complexa. [4]

Colección	**Media** (μ)	**Desv. típ.** (σ)	**Umbral de confianza** ($\tau = \mu + 3\sigma$)
MNIST	0.56	0.23	1.26
MNISTM	0.54	0.39	1.70
SVHN	0.35	0.35	1.40
SYN	0.97	0.39	2.14
USPS	0.40	0.16	0.89

Táboa 3. Métricas estatísticas sobre as que os autoencoders adestrados con cada conxunto basean a súa decisión de modelo a aplicar (valores expresados en 10^{-2}).

4.6. *Estudo por capas*

Para comprender a contribución de cada compoñente de FLProtector, realízase un estudo por capas nos diferentes contornos experimentais. O obxectivo é desenredar as contribucións do deseño de modelo dual e do mecanismo de selección de modelo dinámico en tempo de inferencia. Avaliamos as seguintes tres variantes:

- **Método completo**: O pipeline completo, que inclúe un modelo global robusto aos outliers, un incremento personalizado e un detector de novidade baseado en autoencoders para seleccionar dinámicamente que modelo usar no momento da inferencia.
- **Sen autoencoder**: O mecanismo de detección de novidades é eliminado. No momento da inferencia, cada cliente utiliza sempre o modelo personalizado, independentemente da natureza da mostra de entrada.
- **Sen personalización**: Só se utiliza o modelo global durante a inferencia. Os clientes non calculan nin aplican ningún incremento personalizado. O modelo global segue sendo adestrado utilizando puntuacións de consistencia de gradiente baseadas en L-BFGS para a agregación.
- **Sen ponderación**: O proceso de ponderación baseado en L-BFGS é eliminado. Todos os clientes aportan unha contribución igual ao modelo global, independentemente da súa consistencia ou discrepancia co resto de membros da aprendizaxe.

4.6.1. *Contorno sen ataques*

A táboa 4 mostra o promedio das precisións obtidas polo grupo maioritario de clientes en contorno `d5` (Maioría SVHN) sobre os test local e global en cada variante:

Variante	`d5` (SVHN)		`5b`	
	Local	Global	Local	Global
Completo	82.2 %	70.3 %	89.1 %	70.3 %
Sen autoencod.	82.3 %	60.4 %	89.2 %	55.5 %
Sen personaliz.	72.3 %	72.3 %	78.4 %	78.4 %
Sen ponderac.	81.6 %	73.6 %	89.4 %	69.5 %

Táboa 4. Estudo por capas: precisión media en conxuntos de test local e global. As columnas *Local* reportan a precisión media no conxunto de test propio de cada cliente, medindo a personalización. As columnas *Global* reportan a precisión na unión de todos os conxuntos de test, avaliando a xeneralización ante entradas de varios dominios.

Obsérvase que, en xeral, a versión **completa** do método e aquela **sen ponderación** presentan un rendemento similar en ambos contornos, tanto en termos de precisión local como global. Esta cercanía suxire, a primeira vista, que o mecanismo de ponderación baseado en puntuacións de confianza podería non ser estritamente necesario para garantir un bo comportamento xeral. Por outra banda, a configuración **sen autoencoder** logra unha precisión local igual que a do método completo, mais a costa dunha caída significativa nas precisións globais, poñendo de manifesto a falta de colaboración efectiva entre clientes. No relativo á variante **sen personalización**, é a que consegue obter unha maior precisión global, mais tamén o peor rendemento local, confirmando que un modelo global non pode xeneralizar de maneira adecuada en contornos con distribucións disxuntas.

4.6.2. *Contorno con ataques*

Para retomar a dúbida que xorde da diferencia marxinal entre a versión completa e sen ponderar sobre a utilidade do algoritmo L-BFGS, proponse unha nova batería de probas nun contorno máis adverso, concretamente ante a presencia de ataques maliciosos a un adestramento sobre SVHN, cuxos resultados se presentan na táboa 5. [5]

En primeiro lugar, salienta que cada unha das variantes ofrece un rendemento por separado notablemente superior á referencia de *FedAvg*, evidenciando as capacidades de cada unha para limitar o impacto de posibles discrepancias. Afondando no relativo a cada unha das

[4] A táboa 10 do apéndice B extende esta análise mostrando os resultados de cada autoencoder sobre os conxuntos de test dos outros clientes.

[5] No caso do ataque *Backdoor*, os valores entre parénteses indican o éxito do ataque (imaxes con patrón que foron recoñecidas erroneamente como a clase obxectivo).

Variante	Ataques		
	Backdoor (%)	*Label Flip*	*Mean*
Completo	79.0 % (4 %)	80.3 %	78.4 %
Sen autoencod.	79.0 % (4 %)	80.4 %	78.2 %
Sen personaliz.	84.5 % (14 %)	75.8 %	83.1 %
Sen ponderac.	80.5 % (31 %)	80.5 %	74.5 %
FedAvg	80.4 % (98 %)	69.1 %	10.5 %

Táboa 5. Estudo por capas en contorno de ataque: precisión media dos clientes benignos en test local baixo as diferentes variantes do método proposto e tipos de ataque. Inclúense ademais os resultados do método *FedAvg* como referencia de base.

capas do método, o caso do ataque *backdoor*, onde o patrón malicioso logra infiltrarse naquelas configuracións que prescinden de mecanismos chave do sistema. Concretamente, a variante **sen ponderación** conduce a unha drástica caída da robustez (cunha taxa de acerto do patrón do 31 % fronte ao 4 % da variante completa), demostrando que a ponderación por confianza é esencial para filtrar comportamentos anómalos en contornos con ataques. Por outra banda, a personalización segue aprendendo o patrón malicioso, aínda que en menor medida. A variante **sen autoencoders** mostra practicamente os mesmos resultados que o método completo ao tratarse só de test local, o que reflicte a capacidade do autoencoder de recoñecer patróns coñecidos e empregar sempre o método personalizado. No caso do ataque *mean*, máis agresivo, a configuración completa mantén un rendemento sólido, mentres que a eliminación da ponderación volve a ver deteriorado o seu rendemento.

Un aspecto de especial interese xorde no caso da configuración **sen personalización**, pois ofrece un rendemento salientablemente superior ao do método completo. Este fenómeno suxire que a personalización, se ben eficaz en moitos casos, non sempre resulta beneficiosa cando o dominio local presenta unha elevada complexidade ou ruído (como é o caso de SVHN). Neses escenarios, un modelo global adestrado sobre unha base ampla e compartida pode capturar patróns máis xerais e robustos, evitando sobreaxustes locais. Este feito abre unha liña de investigación futura: o desenvolvemento de estratexias de personalización adaptativa, que permitan a cada cliente regular dinamicamente o grao de personalización aplicado segundo posibles métricas de confianza ou dificultade do dominio local.

4.7. Comparación con métodos do estado da arte e avaliación do rendemento

O obxectivo deste apartado é o de comparar o rendemento de FLProtector fronte diferentes métodos de referencia no estado da arte da aprendizaxe federada baixo as dúas configuracións de heteroxeneidade natural presentadas. Ambos escenarios permiten analizar a capacidade dos métodos para adaptarse á diversidade entre clientes, presentando especial atención á precisión local, é dicir, o rendemento de cada cliente sobre o seu propio conxunto de proba.

Inicialmente, aplícanse análises estatísticas non paramétricas, dado o tamaño reducido da mostra, e céntrase a análise exclusivamente nas observacións de maior interese: **todos os clientes** no caso `5b` (onde todos son igualmente dispares), e o **cliente discrepante** no caso `d5`.

Os métodos considerados nas comparativas son:

- **Individual**: Cada cliente adestra só empregando datos locais, sen ningunha colaboración. Esta liña base reflicte o escenario sen federamento e a dispoñibilidade limitada de datos.
- **FedAvg** [10]: O algoritmo estándar de aprendizaxe federada no que se adestra un único modelo global mediante a media dos modelos actualizados localmente entre os clientes.
- **FedProx** [8]: Unha extensión de *FedAvg* baseada en regularización que incorpora un termo proximal na función de perda local de cada cliente, restrinxindo as actualizacións para que se manteñan cercanas ao modelo global. A intensidade desta regularización contrólase mediante o hiperparámetro μ. Probouse con valores de $\mu \in \{0.01, 0.1, 1\}$ e atopouse que $\mu = 0.01$ proporcionou o mellor rendemento nos escenarios valorados.
- **SCAFFOLD** [6]: Un método de redución de varianza que emprega variables de control para corrixir as actualizacións locais, mitigando a deriva entre clientes en contornos heteroxéneos. Require manter estados de corrección adicionais tanto no servidor como nos clientes, e é implementado habitualmente co optimizador SGD.
- **APFL** [3]: Un método de interpolación explícita de modelos no que cada cliente mantén tanto un modelo global como local. Emprégase un coeficiente de interpolación (α) para combinar ambos modelos durante a inferencia. Os autores tamén propoñen usar SGD para actualizar dinámicamente α durante o adestramento. En concreto, inicializouse cada cliente con $\alpha_0 = 0.5$ e empregouse unha taxa de aprendizaxe de $lr_a = 0.0001$ para actualizar α.
- **FLProtector**: Unha estratexia de dous modelos na que cada cliente aprende un modelo global e un incremento personalizado. En tempo de inferencia, un detector de novidade baseado en autoencoders selecciona se empregar o modelo global ou personalizado. Ademais, a etapa de agregación emprega puntuacións de consistencia de gradiente (estimadas mediante L-BFGS) para ponderar de forma adaptativa a contribución de cada cliente. Adéstrase o incremento personalizado empregando o mesmo número de épocas locais que o modelo principal ($T_i = T_l = 5$) que, como se comentou antes, é consistente en todos os métodos. Ademais, empregouse un autoencoder con 4 capas convolucionais de codificación e 4 capas de decodificación, adestrado durante 4 épocas.

Cabe salientar que tanto FedProx como APFL son altamente sensibles aos seus respectivos hiperparámetros: o coeficiente proximal μ en FedProx, e o factor de interpolación α en APFL. O seu rendemento depende significativamente da elección destes parámetros, que deben

ser axustados coidadosamente para adaptarse á heteroxeneidade dos datos. Nos experimentos, selecciónanse os valores que proporcionan os mellores resultados empíricos para cada método do estado da arte, garantindo unha comparación xusta e competitiva. En contraste, FLProtector non depende de hiperparámetros sensibles para equilibrar a contribución da información local e global. Isto fai que o novo enfoque sexa máis robusto e máis fácil de implementar na práctica. En concreto, dado que todos os clientes teñen o mesmo número de mostras de adestramento na nosa configuración, o factor de ponderación X empregado no esquema de agregación (véxase a ecuación 4) convértese en neutro ($X = 0$), eliminando a necesidade de axustes adicionais.

Agárdase que FLProtector, ao incorporar un sistema de selección personalizada por mostra e unha ponderación de contribucións baseada en confianza supere en precisión local aos métodos federados clásicos (FedAvg, FedProx, SCAFFOLD) e se sitúe en niveis comparables a APFL, método de referencia na aprendizaxe federada personalizada. Para verificar esta hipótese aplicouse o test de Friedman, considerando os cinco clientes como bloques emparellados. Os resultados, mostrados na táboa 6 permiten en ambos casos rexeitar a hipótese nula de igualdade entre métodos, indicando diferencias estatísticamente significativas na precisión local. Como análise complementaria, aplicouse o test de Wilcoxon pareado entre FLProtector e cada un dos métodos do estado da arte.

Test	**Comparación**	`5b`	`d5`
Friedman	-	$H = 17.49$	$H = 14.37$
	-	$p = 0.0015$	$p = 0.0062$
Wilcoxon	vs APFL	0.125	0.695
	vs FedAvg	**0.0625**	**0.0195**
	vs FedProx	**0.0625**	**0.0137**
	vs Scaffold	**0.0625**	**0.0284**

Táboa 6. Resultados dos tests de Friedman e Wilcoxon (FLProtector fronte o resto) para os dous contornos: `5b` (todos os clientes) e `d5` (cliente discrepante).

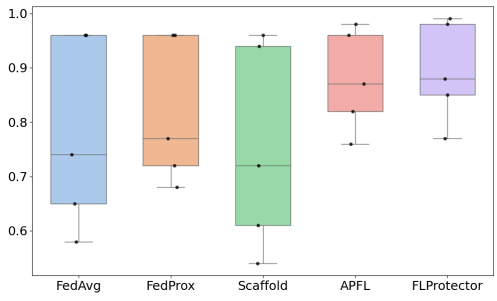

Figura 9. Distribución das precisións locais dos cinco clientes no contorno `5b` segundo o método de federamento empregado.

Aínda que os valores de p non acadan significancia estrita ($p < 0{,}05$), no contorno `5b` debido ao número reducido de mostras ($n = 5$), todos apuntan cara unha vantaxe de FLProtector sobre os métodos clásicos que, ademais, se fai visible na figura 9. Por outra banda, no contorno `d5`, os valores obtidos si confirman que FLProtector mellora significativamente a precisión local do cliente discrepante fronte aos métodos referencia de agregación. Fronte a APFL non se detectan diferencias significativas, o que posiciona a proposta deste traballo ao nivel do estado de arte na aprendizaxe federada personalizada.

A táboa 7 presenta os resultados obtidos polos métodos avaliados en contornos de heteroxeneidade parcial (`d5`) e completa (`5b`), tanto en test local como global. A columna *Promedio* mostra a precisión media obtida por cada método en todos os contornos. Cada cela da táboa representa a precisión media sobre 5 execucións, unha por cada colección de datos. Por exemplo, os valores mostrados na columna *MNIST* (Discr.)* corresponden á media de rendemento dos clientes non-MNIST en todas as execucións nas que MNIST foi o dominio maioritario. É dicir, inclúe o rendemento dun cliente MNIST-M adestrado xunto a 4 clientes MNIST, un cliente SVHN con 4 clientes MNIST, un cliente USPS con 4 clientes MNIST e un cliente SYN con 4 clientes MNIST. De maneira similar, a columna *MNIST (Maioría)* informa da precisión media dos catro clientes maioritarios de MNIST nesas mesmas probas.

Cando a maioría de clientes pertencen a un dominio sinxelo como MNIST, pódese observar na primeira columna que o cliente discrepante (que adestra con MNIST-M, USPS, SYN ou SVHN) logra unha precisión local significativamente inferior en comparación cos clientes maioritarios (segunda columna). Isto é esperable, xa que o cliente discrepante non se beneficia significativamente da colaboración. Curiosamente, os únicos métodos que amosan unha mellora de rendemento nos test locais grazas á colaboración—en comparación co adestramento individual— son APFL e FLProtector. Unha tendencia similar obsérvase cando a maioría de clientes pertence a SVHN.

Á hora de analizar as precisións obtidas pola maioría dos clientes, obsérvase que os rendementos sobre test local son consistentemente maiores que o dos casos *individual*, feito que reflicte o beneficio da colaboración entre os clientes e pon en valor a aplicación da aprendizaxe federada. Resulta interesante que o rendemento acadado mediante FedAvg, pese á súa simpleza algorítmica, sexa dos máis elevados, o que suxire que a presenza dun único cliente diferente (fronte aos 4 clientes do mesmo tipo) non prexudica gravemente ao modelo global.

No relativo ao caso `5b`, onde cada cliente adestra con datos diferentes, obsérvase que tan só APFL e FLProtector son capaces de acadar mellores resultados que o adestramento individual sobre os test locais, o que indica que estes métodos realmente se benefician da colaboración en contornas de alta heteroxeneidade. Pola contra, os métodos que non logran un bo rendemento nos test locais son capaces de xeneralizar relativamente ben nos test globais. Isto débese a que estes métodos producen modelos que intentan adaptarse a todos os conxuntos de datos, resultando nunha boa xenerali-

TEST LOCAL	d5				5b	Promedio
	MNIST* (Discr.)	MNIST (Maioría)	SVHN* (Discr.)	SVHN (Maioría)	5 Conxuntos	
Individual	87.4 %	97.0 %	87.4 %	77.0 %	87.8 %	87.3 %
FedAvg	75.2 %	98.4 %	75.3 %	82.5 %	78.1 %	81.9 %
FedProx	74.9 %	98.2 %	73.6 %	83.0 %	75.2 %	81.0 %
Scaffold	83.8 %	97.5 %	78.4 %	81.9 %	81.6 %	84.6 %
APFL	88.0 %	98.0 %	89.1 %	81.6 %	88.3 %	89.0 %
FLProtector	89.3 %	98.0 %	87.8 %	80.3 %	89.1 %	88.9 %

TEST GLOBAL	d5				5b	Promedio
	MNIST* (Discr.)	MNIST (Maioría)	SVHN* (Discr.)	SVHN (Maioría)	5 Conxuntos	
Individual	76.8 %	71.6 %	58.8 %	59.6 %	45.5 %	62.5 %
FedAvg	86.8 %	86.8 %	78.9 %	78.9 %	78.1 %	81.9 %
FedProx	86.5 %	86.5 %	78.3 %	78.3 %	75.2 %	81.0 %
Scaffold	90.6 %	90.6 %	80.1 %	80.1 %	81.6 %	84.6 %
APFL	82.4 %	78.4 %	63.5 %	70.0 %	54.9 %	69.8 %
FLProtector	88.5 %	87.8 %	75.2 %	71.6 %	70.3 %	78.5 %

Táboa 7. Precisións sobre test local e global obtidas por cada un dos métodos nos diferentes contornos de heteroxeneidade de datos: `d5` (un cliente discrepante respecto os 4 clientes que comparten conxunto fácil (MNIST) ou difícil (SVHN)) e `5b` (5 clientes con diferentes coleccións). Os resultados de `d5` son a media de 4 execucións, unha por cliente discrepante diferente. NOTA: Os resultados detallados de cada unha das execucións adxúntanse na táboa 11 no apéndice final.

zación pero cunha menor precisión en calquera conxunto de datos individual. É importante destacar que APFL amosa un rendemento deficiente nos test globais, o que suxire que o valor de α fai que o modelo empregado por cada cliente sexa practicamente local. En canto a FLProtector, aínda que os resultados globais non acadan a mesma alta precisión que outros métodos, é capaz de xeneralizar relativamente ben, acadando arredor dun 70 % de precisión, ao tempo que logra consistentemente os mellores resultados locais.

Métodos como FedProx e SCAFFOLD, que perseguen aprender un único modelo global capaz de xeneralizar entre todos os clientes, logran un bo rendemento nos test globais e para a maioría dos clientes, mais co prezo de descoidar as necesidades dos clientes minoritarios. En contraste, APFL busca un equilibrio entre modelos locais e globais, o que lle permite adaptarse ben aos test locais, porén o seu rendemento nas avaliacións globais segue sendo limitado. FLProtector, pola súa banda, logra un mellor equilibrio: pode personalizarse de forma efectiva para os clientes minoritarios, apoiar a aprendizaxe colaborativa entre os clientes maioritarios e manter unha boa xeneralización mediante o uso do mecanismo de decisión baseado en autoencoders.

5. Conclusións e traballo futuro

Este traballo propón **FLProtector**, un novo enfoque de aprendizaxe federada personalizada que combina dous mecanismos complementarios:

- Un modelo dual formado por unha base global e un incremento personalizado local
- Un selector dinámico baseado en autoencoders que determina, en tempo de inferencia, que modelo empregar segundo a distribución da mostra de entrada.

Os experimentos realizados sobre o *benchmark Digit-Five* demostran que FLProtector consegue un equilibrio robusto entre xeneralización e personalización, superando de forma consistente tanto a métodos do estado da arte clásicos (FedAvg, FedProx e SCAFFOLD) como a alternativas personalizadas (APFL). A arquitectura proposta amosa unha especial capacidade para adaptarse a clientes discrepantes sen comprometer o rendemento do resto do federamento. Ademais, intégrase un mecanismo de ponderación por consistencia, inspirado en métodos do estado da arte de detección de clientes maliciosos, que mellora a robustez do agregado global en situacións de alta heteroxeneidade ou comportamento malicioso.

A análise realizada identifica tamén dúas limitacións importantes, que abren novas liñas de investigación:

- En primeiro lugar, a **elección do modelo a empregar por parte do autoencoder non é igualmente precisa en todos os dominios**. Isto suxire a necesidade de mecanismos adaptativos de selección, que pode pasar por autoencoders máis especializados ou alternativas que permitan ter en conta as características específicas de cada dominio.
- En segundo lugar, a **personalización non é sempre beneficiosa**, como se observa en dominios difíciles onde o modelo global pode capturar mellor os patróns xerais. Isto apunta a una liña de investigación na que se explore a regularización dinámica do grao de personalización segundo o nivel de confianza ou dificultade da tarefa, o que permitiría un axuste fino entre especialización e xeneralización.

Finalmente, como continuación natural do traballo, proponse avaliar FLProtector en tarefas de maior complexidade. En particular, proponse a extensión ao ámbito da robótica, onde diferentes robots ou sensores poden

manexar datos similares pero adquiridos con diferentes condicións. Xa en traballos previos [9] se observou que as puntuacións de confianza en tarefas do estilo resultaban ser prometedoras para detectar e modular esas diferenzas, o que da pé á aplicabilidade do enfoque máis aló de conxuntos de datos sintéticos.

Agradecementos

A quen, *polos seus actos*, os mereza.

Referencias

[1] R. H. Byrd, J.Ñocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1): 129–156, 1994.

[2] X. Cao, J. Jia, Z. Zhang, and N. Z. Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information, 2022. URL https://arxiv.org/abs/2210.10936.

[3] Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. Preprint arXiv:2003.13461, 2020.

[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks, 2016.

[5] J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

[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, 2021. URL https://arxiv.org/abs/1910.06378.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[8] T. Li, A. Kumar Sahu, M. Zahee, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. In *Proceedings of the 3rd MLSys Conference*, pages 429–450, 2020.

[9] R. Martínez Enríquez, R. Iglesias Rodríguez, and S. Barro Ameneiro. Detección de clientes maliciosos ou ruidosos na aprendizaxe federada. https://minerva.usc.gal/entities/publication/5b299596-a6de-4b2f-8641-a87eb27751bf, 2024. Trabajo Fin de Grao, Universidade de Santiago de Compostela.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

[11] Y.Ñetzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[12] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal. Effects of degradations on deep neural network architectures, 2023.

[13] L. Seinmann, N. Migenda, T. Voigt, M. Kohlhase, and W. Schenck. Variational autoencoder based novelty detection for real-world time series. In *MSIE '21: Proceedings of the 2021 3rd International Conference on Management Science and Industrial Engineering*, pages 1–7, 2021.

[14] Z. Zhang, X. Cao, J. Jia, and N. Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2545–2555, 2022.

[15] A. Ziying Tan, H. Yu, L. Cui, and Q. Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023.

A. Arquitectura dos modelos e especificación do hardware

Este apéndice proporciona un desglose das arquitecturas das redes neuronais empregadas nos experimentos descritos neste traballo, así como as especificacións do hardware no que as probas foron realizadas.

Rede	Capa	Tamaño
DigitFive	Entrada	3 x 16 x 16
	Convolución + ReLU	3 x 16 x 16 (16 filtros 3x3)
	Max Pooling	2 x 2
	Convolución + ReLU	16 x 16 x 16 (32 filtros 3x3)
	Max Pooling	2 x 2
	Convolución + ReLU	32 x 16 x 16 (32 filtros 3x3)
	Max Pooling	2 x 2
	Convolución + ReLU	32 x 16 x 16 (32 filtros 3x3)
	Max Pooling	2 x 2
	Fully Connected + ReLU	$32 \rightarrow 64$
	Fully Connected + ReLU	$64 \rightarrow 32$
	Fully Connected + ReLU	$32 \rightarrow 16$
	Fully Connected	$16 \rightarrow 10$
Autoencoder	Entrada	3 x 16 x 16
	Convolución + ReLU	16 x 16 x 16 (16 filtros 3x3)
	Max Pooling	8 x 8 x 16
	Convolución + ReLU	8 x 8 x 8 (8 filtros 3x3)
	Max Pooling	4 x 4 x 8
	ConvTranspose + ReLU	8 x 8 x 16
	ConvTranspose + Sigmoid	16 x 16 x 3

Táboa 8. Arquitectura das redes empregadas: **DigitFive** para clasificación e **Autoencoder** para reconstrución.

Compoñente	Especificación
Modelo	Micro-Star International Co., Ltd. Pulse GL76 12UEK
CPU	Intel Core i7-12700HQ @ 2.70GHz – 14 procesadores
GPU	NVIDIA GeForce RTX 3060 Laptop (6GB)
RAM	32 GB
SO	Microsoft Windows 11 Pro v10.0.22631

Táboa 9. Especificacións do hardware empregado para os experimentos.

B. Resultados extendidos da análise experimental

Adestramento	**Test**				
	MNIST	**MNISTM**	**SVHN**	**SYN**	**USPS**
MNIST	0.56 ± 0.24	5.30 ± 2.31	5.42 ± 1.86	6.44 ± 1.76	1.84 ± 1.06
MNISTM	1.18 ± 0.15	0.71 ± 0.65	0.29 ± 0.22	1.79 ± 0.97	1.32 ± 0.22
SVHN	2.92 ± 0.18	1.11 ± 1.01	0.35 ± 0.33	2.35 ± 1.19	3.04 ± 0.37
SYN	1.56 ± 0.16	0.48 ± 0.32	0.20 ± 0.15	1.06 ± 0.48	2.08 ± 0.33
USPS	0.29 ± 0.11	4.20 ± 1.68	4.41 ± 1.26	4.74 ± 1.54	0.34 ± 0.13

Táboa 10. Resultados de avaliación cruzada entre autoencoders adestrados sobre o dominio local (filas) e avaliados sobre imaxes de todos os dominios (columnas). Cada celda mostra a perda de reconstrucción media ± desviación estándar (valores expresados en 10^{-2}).

d5: **Maioría MNIST**								
TEST LOCAL	MNISTM	*Mai.*	SVHN	*Mai.*	SYN	*Mai.*	USPS	*Mai.*
Individual	81.0 %	97.0 %	75.0 %	97.0 %	85.0 %	97.0 %	99.0 %	97.0 %
FedAvg	70.7 %	98.2 %	46.3 %	98.5 %	63.0 %	98.4 %	97.6 %	98.1 %
Scaffold	79.2 %	97.3 %	70.8 %	97.0 %	72.9 %	97.4 %	97.7 %	97.6 %
FedProx	71.7 %	98.3 %	49.2 %	98.1 %	57.6 %	98.3 %	97.4 %	98.0 %
APFL	82.3 %	98.1 %	76.8 %	98.1 %	84.3 %	98.0 %	98.8 %	97.9 %
FLProte.	85.4 %	97.8 %	77.3 %	98.1 %	86.9 %	98.0 %	98.7 %	98.0 %
TEST GLOBAL	MNISTM	*Mai.*	SVHN	*Mai.*	SYN	*Mai.*	USPS	*Mai.*
Individual	86.0 %	62.0 %	61.0 %	53.0 %	56.0 %	56.0 %	84.0 %	90.0 %
FedAvg	84.5 %	84.5 %	72.4 %	72.4 %	80.7 %	80.7 %	97.8 %	97.8 %
Scaffold	88.2 %	88.2 %	83.9 %	83.9 %	85.1 %	85.1 %	97.6 %	97.6 %
FedProx	85.0 %	85.0 %	73.6 %	73.6 %	77.9 %	77.9 %	97.7 %	97.7 %
APFL	88.8 %	73.9 %	64.6 %	60.6 %	64.1 %	63.2 %	96.5 %	96.0 %
FLProte.	90.2 %	86.6 %	87.9 %	78.2 %	70.3 %	79.1 %	96.0 %	96.8 %

d5: **Maioría SVHN**								
TEST LOCAL	MNIST	*Mai.*	MNISTM	*Mai.*	SYN	*Mai.*	USPS	*Mai.*
Individual	97.0 %	97.0 %	81.0 %	97.0 %	85.0 %	97.0 %	99.0 %	97.0 %
FedAvg	92.0 %	80.9 %	57.2 %	82.0 %	52.9 %	81.8 %	89.4 %	82.9 %
Scaffold	93.8 %	82.7 %	66.6 %	81.6 %	54.0 %	80.8 %	95.2 %	81.7 %
FedProx	87.1 %	82.0 %	58.3 %	82.6 %	49.7 %	82.3 %	87.9 %	82.9 %
APFL	97.0 %	81.3 %	82.9 %	82.0 %	84.1 %	89.8 %	98.0 %	81.1 %
FLProte.	97.1 %	80.2 %	80.8 %	80.5 %	85.8 %	80.0 %	98.1 %	79.7 %
TEST GLOBAL	MNIST	*Mai.*	MNISTM	*Mai.*	SYN	*Mai.*	USPS	*Mai.*
Individual	56.0 %	58.0 %	55.0 %	52.0 %	53.0 %	47.0 %	55.0 %	64.0 %
FedAvg	86.5 %	86.5 %	69.6 %	69.6 %	67.3 %	67.3 %	86.2 %	86.2 %
Scaffold	88.2 %	88.2 %	74.1 %	74.1 %	67.4 %	67.4 %	88.4 %	88.4 %
FedProx	84.5 %	84.5 %	70.4 %	70.4 %	66.0 %	66.0 %	85.4 %	85.4 %
APFL	59.5 %	73.2 %	61.6 %	62.0 %	56.8 %	54.5 %	56.5 %	77.8 %
FLProtec.	90.5 %	81.1 %	60.9 %	59.5 %	56.6 %	55.7 %	90.8 %	80.6 %

Táboa 11. Resultados detallados sobre test local e global de cada unha das 5 execucións d5 coa maioría de clientes traballando con MNIST ou SVHN, mentres que o cliente discrepante cambia de conxunto en cada execución. *Mai* mostra a precisión media dos 4 clientes.

5b					
TEST LOCAL	MNIST	MNISTM	SVHN	SYN	USPS
Individual	97.2 %	81.5 %	77.3 %	84.3 %	98.7 %
FedAvg	96.4 %	74.3 %	58.4 %	65.3 %	96.2 %
FedProx	96.3 %	71.6 %	53.5 %	60.8 %	94.0 %
Scaffold	96.1 %	76.5 %	67.8 %	69.5 %	96.1 %
APFL	97.7 %	82.4 %	76.4 %	76.7 %	98.3 %
FLProtector	97.6 %	84.5 %	77.1 %	87.9 %	98.7 %
TEST GLOBAL	MNIST	MNISTM	SVHN	SYN	USPS
Individual	47.6 %	54.4 %	41.7 %	39.1 %	44.8 %
FedAvg	78.1 %	78.1 %	70.1 %	68.1 %	78.1 %
FedProx	75.2 %	75.2 %	75.2 %	70.4 %	78.1 %
Scaffold	81.6 %	81.6 %	81.6 %	81.6 %	81.6 %
APFL	58.0 %	65.6 %	45.9 %	55.3 %	67.0 %
FLProtector	78.4 %	66.3 %	75.8 %	54.2 %	76.6 %

Táboa 12. Resultados en test local e global no contorno 5b, no que cada cliente traballa cun conxunto diferente.