Inteligencia Artificial Explicable para Resolver la Heterogeneidad en el Aprendizaje Federado

Mario García-Márquez[1], Nuria Rodríguez-Barroso[1], Francisco Herrera[1], and M.V. Luzón[2]

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España
Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI)
{mariogmarq, rbnuria}@ugr.es, herrera@decsai.ugr.es

[2] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Granada, Granada, España
Instituto Andaluz Interuniversitario en Ciencia de Datos e Inteligencia Computacional (DaSCI)
luzon@ugr.es

Resumen El Aprendizaje Federado permite el entrenamiento colaborativo de modelos sin compartir datos, convirtiéndolo en una solución prometedora para aplicaciones sensibles a la privacidad. Sin embargo, su rendimiento a menudo se degrada debido a datos no-IID entre los clientes, lo que provoca una deriva del modelo y fronteras de decisión inconsistentes. Planteamos la hipótesis de que esta degradación está causada en parte por la desalineación de las características aprendidas por los diferentes clientes. Para abordar este desafío, proponemos la Ponderación de Relevancia de Conceptos Federada, FedPRC, un mecanismo de agregación novedoso que aprovecha la Propagación de Relevancia de Conceptos para ponderar las contribuciones de los clientes, basándose en la relevancia de sus características aprendidas. Al aprovechar un pequeño conjunto de validación en el servidor para identificar las características que son consistentemente relevantes entre los clientes, FedPRC mitiga eficazmente los efectos adversos de la heterogeneidad de los datos, produciendo representaciones globales más robustas y alineadas. Los resultados experimentales muestran que FedPRC iguala o supera a los métodos de optimización federados del estado del arte, a la vez que proporciona una mayor robustez frente a la heterogeneidad, una alineación de características más fuerte y una mejor interpretabilidad del modelo.

Keywords: Aprendizaje Federado · No-IID · IA Explicable · Redes Heterogéneas · Relevancia de Conceptos

1. Introducción

El rápido desarrollo de la Inteligencia Artificial (IA) ha revolucionado numerosos sectores, desde la salud hasta las finanzas, permitiendo niveles sin precedentes de automatización y capacidad de toma de decisiones [19]. Sin embargo,

este progreso ha generado una creciente preocupación en torno a la privacidad de los datos y la concentración de información sensible en repositorios centralizados. Estas preocupaciones han impulsado marcos regulatorios, como el Reglamento General de Protección de Datos (GDPR) [1] y la Ley de IA (AI Act) [9] en Europa, y regulaciones similares en todo el mundo [21], a imponer restricciones estrictas sobre cómo se pueden recopilar, almacenar y procesar los datos sensibles. En respuesta a estas presiones regulatorias y a las limitaciones inherentes de los paradigmas de aprendizaje centralizado, el Aprendizaje Federado (AF) [18] ha surgido como un enfoque destacado que permite a múltiples partes entrenar de manera colaborativa modelos de aprendizaje automático sin transferir sus datos a una autoridad central.

A pesar de sus ventajas para preservar la privacidad, el AF se enfrenta a un desafío crítico: los datos que poseen los clientes individuales suelen ser no-IID [12,5]. Esta heterogeneidad se manifiesta de múltiples formas, incluyendo diferencias en las distribuciones de etiquetas entre clientes, variaciones en los espacios de características y cantidades desiguales de datos, lo que provoca una degradación significativa del rendimiento en comparación con los escenarios de aprendizaje centralizado. Los algoritmos de optimización tradicionales diseñados bajo la hipótesis IID, como FedAvg [20], con frecuencia no logran converger a soluciones satisfactorias en entornos heterogéneos, mostrando en su lugar puntos de convergencia subóptimos o inconsistentes [16]. Aunque se han propuesto algoritmos especializados, como FedProx [16], FedNova [23] y FedDyn [2], para mitigar este desafío mediante diversas estrategias de regularización y normalización, el problema fundamental de la deriva del modelo (model drift) sigue estando en gran medida sin resolver, especialmente en escenarios de heterogeneidad extrema.

Planteamos la hipótesis de que gran parte de la degradación del rendimiento en el AF heterogéneo se debe a la desalineación de las características aprendidas entre los clientes. Específicamente, mientras que los extractores de características aprenden de forma natural representaciones diversas adaptadas a las distribuciones de datos locales [8], las cabezas de clasificación resultantes se desalinean cada vez más, lo que genera fronteras de decisión inconsistentes en toda la red. Si los clientes pudieran alinear sus características aprendidas con una comprensión común de qué características son universalmente relevantes, conjeturamos que la generalización mejoraría, ya que el modelo global exhibiría una mayor robustez frente a los cambios en la distribución local.

Para validar esta hipótesis y abordar el desafío de la heterogeneidad, proponemos la **P**onderación de **R**elevancia de **C**onceptos **Fed**erada (**FedPRC**), una propuesta novedosa que aprovecha la relevancia de los conceptos para guiar la agregación federada. Más precisamente, la innovación central de FedPRC es el uso de la herramienta de Propopagación de Relevancia de Conceptos (CRP) [3], una técnica de Inteligencia Artificial Explicable (XAI) que descompone las predicciones del modelo para cuantificar la importancia de las características latentes individuales, para ponderar las contribuciones de cada cliente durante la agregación de la cabeza de clasificación. Al anclar las decisiones de agregación a

un pequeño conjunto de datos de validación en el servidor y calcular puntuaciones de relevancia para el clasificador de cada cliente, FedPRC promueve el uso de características que son consistentemente útiles en toda la red, a la vez que suprime las características específicas de cada cliente. Este enfoque transforma una herramienta de interpretabilidad post-hoc en una parte activa del proceso de aprendizaje.

Para evaluar rigurosamente FedPRC, diseñamos nuestro análisis experimental para responder a cuatro preguntas de investigación centrales: cómo se compara la propuesta con los estados del arte bajo diferentes tipos de heterogeneidad (PI1), su resiliencia ante sesgos mixtos complejos (PI2), su robustez bajo desequilibrios extremos (PI3), y su capacidad para mejorar la alineación de características y la interpretabilidad (PI4). Con el fin de responder a estas preguntas, llevamos a cabo experimentos exhaustivos empleando CIFAR-10 con un sesgo de etiquetas extremo para estresar los escenarios de desequilibrio, y la suite de conjuntos de datos CelebA (Smiling, Makeup y Attractive), que presenta mezclas realistas de sesgo de etiqueta, características y cantidad. Comparamos FedPRC con un conjunto diverso de métodos de referencia que representan el estado del arte en técnicas de optimización federada: FedAvg [20] como el método base fundacional, FedProx [16] y FedNova [23] como métodos destacados sensibles a la heterogeneidad, y FedDyn [2] como un enfoque reciente de regularización dinámica. Nuestros resultados demuestran que FedPRC logra un rendimiento altamente competitivo en términos de precisión y pérdida en todos los escenarios y, fundamentalmente, produce modelos globales más significativos desde el punto de vista semántico e interpretables.

El resto de este trabajo se estructura de la siguiente manera. La Sección 2 detalla los antecedentes esenciales de AF, no-IID y CRP. La Sección 3 introduce la propuesta FedPRC. La Sección 4 describe el entorno experimental. La Sección 5 presenta y analiza los resultados empíricos obtenidos. Finalmente, la Sección 6 expone las conclusiones, limitaciones y trabajo futuro.

2. Antecedentes

Esta sección describe los conceptos fundamentales del AF, la heterogeneidad estadística de los datos (no-IID) y de CRP.

El AF entrena colaborativamente un modelo global $\mathcal{G}$ entre n clientes sin transferir datos [20]. En una configuración síncrona, cada ronda de aprendizaje t implica que los clientes optimizan los modelos locales $L_i^t = g \circ f$ (donde f es el extractor de características y g la cabeza de clasificación) utilizando conjuntos de datos privados D_i. El servidor agrega los pesos actualizados V_i^t a través de una regla de agregación Δ (por ejemplo, FedAvg) para producir $V_{\mathcal{G}}^{t+1} = \Delta(V_1^t, \ldots, V_n^t)$, que se devuelve a los clientes [18]. El AF funciona a través de dos etapas distintas: (1) una Etapa de Entrenamiento, donde cada cliente entrena un modelo local con datos privados y envía sus pesos al servidor, y (2) una Etapa de Inferencia, donde los clientes realizan predicciones de forma cooperativa utilizando el modelo global entrenado.

En escenarios del mundo real, las distribuciones de datos de los clientes $\mathbf{P}_i(x, y)$ suelen ser no-IID ($\mathbf{P}_i(x, y) \neq \mathbf{P}_j(x, y)$), lo que degrada la convergencia de la agregación estándar. La heterogeneidad se manifiesta de tres formas [12]: (1) sesgo de etiquetas ($\mathbf{P}_i(y) \neq \mathbf{P}_j(y)$), (2) sesgo de características ($\mathbf{P}_i(x) \neq \mathbf{P}_j(x)$) y (3) sesgo de cantidad ($|D_i| \neq |D_j|$). Aunque los algoritmos del estado del arte abordan la heterogeneidad estadística (por ejemplo, FedProx [16] mediante restricciones proximales de proximidad, FedNova [23] mediante actualizaciones normalizadas de los clientes y FedDyn [2] mediante regularización dinámica), la deriva del modelo de los clientes sigue siendo un cuello de botella persistente bajo sesgos estadísticos extremos.

Para abordar esta deriva, nuestra propuesta integra conceptos de XAI. Concretamente, CRP [3]. CRP cierra la brecha entre las atribuciones locales a nivel de píxel y la semántica global de los conceptos. Extendiendo la Propagación de Relevancia por Capas [4], que distribuye la salida de una predicción $f(x)$ hasta obtener puntuaciones de relevancia de entrada, CRP aísla y rastrea los conceptos latentes de los canales. Al condicionar el paso hacia atrás (backward pass) en filtros de capas latentes específicos, CRP calcula mapas de relevancia condicional $R(x|\theta)$ que cuantifican la importancia de representaciones latentes específicas para la decisión final.

3. FedPRC: Ponderación de Relevancia de Conceptos Federada

En esta sección presentamos nuestra propuesta: FedRPC. Intuitivamente, FedPRC intenta incentivar a los clientes a utilizar características que sean útiles globalmente, a la vez que evita el uso de características específicas que solo son útiles para un cliente específico, con el fin de lograr una mejor generalización. Esta idea se basa en hallazgos empíricos que muestran que en redes heterogéneas, mientras que los extractores de características (f) aprenden representaciones diversas, la desalineación a menudo se manifiesta de manera más severa en las fronteras de decisión (g) [14]. Por lo tanto, permitimos que f se agregue de manera natural, pero imponemos restricciones estrictas basadas en la relevancia sobre g para alinear la frontera de decisión del modelo global.

Para que FedPRC pueda funcionar, requiere de dos condiciones principales: (1) que el servidor disponga de un pequeño conjunto de datos etiquetado D_S, que contiene una representación mínima de todas las clases (siendo este mantenido de forma privada en el servidor para evitar filtraciones), y (2) que los modelos locales se descompongan como $g \circ f$, donde g es una cabeza de clasificación lineal, tal que $g(x) = [\langle g_1, x\rangle, \ldots, \langle g_{|etiquetas|}, x\rangle]$ para $x, g_i \in \mathbb{R}^z$.

En cada ronda, cuando los parámetros locales se envían al servidor, podemos recuperar los modelos locales (ya que están parametrizados por sus parámetros V_i). El mecanismo FedPRC propuesto consiste en una agregación ponderada para la última capa g que potencia las entradas con alta relevancia para todos los clientes. Para un g_i dado, la agregación ponderada se define por la relevancia normalizada en la Ec. 1, donde g_{ij} es la j-ésima columna para el i-ésimo cliente,

que también corresponde al límite de decisión para la j-ésima clase para dicho cliente.

$$\tilde{R}(g_{ij}) = \frac{|R(g_{ij}|\theta)| - \min|R(g_{ij}|\theta)|}{\max|R(g_{ij}|\theta)| - \min|R(g_{ij}|\theta)|} \tag{1}$$

Además, se desearía prestar más atención a aquellos clientes que logran un mejor rendimiento. Para ello, calculamos la relevancia ponderada de un cliente, que es un escalar, multiplicando su relevancia normalizada por la probabilidad de clasificar correctamente la muestra, como se muestra en la Ec. 2.

$$w_{ij} = \frac{e^{Pred_{ij}}}{\sum_k^n e^{Pred_{ik}}} \cdot \tilde{R}_{ij} \tag{2}$$

Finalmente, definimos los parámetros de la cabeza del modelo global agregado como la suma ponderada de las diferentes fronteras de decisión de los clientes. El procedimiento para un $g_{.j}$ dado se muestra en la Ec. 3.

$$V_{\mathcal{G},g_j}^{t+1} = \sum_{i \in S_t} w_{ij} \cdot V_{i,g_j}^{t} \tag{3}$$

Para el resto de la red, simplemente aplicamos FedAvg como se muestra en la Ec. 4.

$$V_{\mathcal{G},f}^{t+1} = \frac{1}{|S_t|} \sum_{j \in S_t} V_{j,f}^{t} \tag{4}$$

Al escalar la relevancia y aplicar una función softmax a las predicciones, se garantiza que w_{ij} esté en el intervalo $[0, 1]$. De este modo, cada frontera del modelo global es de hecho una combinación convexa de las fronteras de decisión de los clientes, lo que permite la estabilidad numérica. De esta manera, FedPRC produce una alineación de las fronteras de decisión. Al aprovechar el conjunto de datos de anclaje del servidor D_S como un validador de referencia (ground-truth) para la relevancia de las características, suprimimos dinámicamente los sesgos específicos de los clientes que suelen surgir en la capa de clasificación final debido a la heterogeneidad de los datos. En consecuencia, el modelo global no se construye simplemente promediando parámetros diversos, sino integrando selectivamente solo aquellos componentes de decisión que demuestran una relevancia de conceptos robusta, consistente y aplicable, mejorando así significativamente el rendimiento de generalización en entornos no-IID sin comprometer la privacidad de los clientes locales.

4. Entorno Experimental

En esta sección, describimos el entorno experimental utilizado para evaluar FedPRC. Explicamos los conjuntos de datos y modelos empleados en la Sección 4.1, definimos los métodos de referencia del estado del arte para la comparación en la Sección 4.2, y proporcionamos los detalles de implementación necesarios para la reproducibilidad en la Sección 4.3.

4.1. Conjuntos de Datos y Modelos

Para la evaluación, empleamos cuatro bancos de pruebas de clasificación de imágenes distribuidos entre los clientes para reflejar entornos no-IID realistas: (1) **CIFAR-10** [15], particionado entre 500 clientes mediante una distribución de Dirichlet ($\alpha = 0.1$) para simular un sesgo de etiquetas severo [23]; (2) **CelebA-S** [17], configurado como una tarea de clasificación binaria para el atributo *Smiling* (Sonriente) particionado por el ID de la celebridad (filtrando clientes con > 30 muestras, dando como resultado 1,800 clientes), creando una mezcla compleja de sesgos de etiquetas, características y cantidad; (3) **CelebA-M**, utilizando la misma partición dirigida al atributo *Heavy Makeup* (Maquillaje); y (4) **CelebA-A**, dirigido al atributo subjetivo *Attractive* (Atractivo). Para todos los experimentos, el conjunto del servidor D_S consiste en un mínimo de 3 muestras por clase seleccionadas al azar y eliminadas del conjunto de prueba para garantizar que no haya filtración de datos.

4.2. Métodos de Referencia

Para validar el rendimiento de FedPRC, lo comparamos con varios algoritmos del estado del arte diseñados para mitigar la heterogeneidad de los datos: (1) **FedAvg** [20] como base fundacional; (2) **FedProx** [16], que introduce regularización proximal para limitar la deriva local; (3) **FedNova** [23], que normaliza las actualizaciones para corregir inconsistencias del objetivo local; y (4) **FedDyn** [2], que emplea penalizaciones dinámicas para alinear asintóticamente los objetivos locales con el global.

4.3. Detalles de Implementación

Los experimentos están implementados en la librería FLEXible [11][3] con PyTorch [22], empleando una red `ResNet-18` [10] entrenada desde cero. El entrenamiento local utiliza el optimizador Adam (tasa de aprendizaje 0.001) durante 5 épocas por ronda, para un total de 100 rondas. Los clientes se seleccionan aleatoriamente en cada ronda. Reportamos la media y desviación estándar de la precisión y pérdida sobre 5 ejecuciones con semillas aleatorias distintas.

5. Resultados y Análisis

En esta sección, presentamos una evaluación cuantitativa y cualitativa de FedPRC. Primero formalizamos los objetivos de nuestro análisis a través de un conjunto de preguntas de investigación en la Sección 5.1, luego discutimos los resultados empíricos obtenidos en varios entornos heterogéneos en la Sección 5.2. Finalmente, señalamos algunas limitaciones de nuestra propuesta en la Sección 5.3.

[3] https://github.com/FLEXible-FL/FLEXible

5.1. Preguntas de Investigación

Para evaluar rigurosamente la efectividad de FedPRC, nuestro análisis busca responder a cuatro preguntas de investigación clave: **Rendimiento Comparativo (PI1):** ¿Cómo se compara FedPRC con los métodos de referencia del estado del arte bajo sesgos de etiquetas, características y cantidad? **Resiliencia ante sesgos mixto (PI2):** ¿Ofrece la ponderación basada en la relevancia una ventaja en escenarios complejos de sesgos mixtos? **Robustez ante desequilibrios extremos (PI3):** ¿Mantiene FedPRC la estabilidad bajo un sesgo de etiquetas extremo donde otros métodos regularizados fallan? **Alineación de características e interpretabilidad (PI4):** ¿Genera la agregación ponderada por relevancia características globales mejor alineadas y semánticamente significativas?

5.2. Análisis Experimental

Esta sección proporciona una evaluación cuantitativa de los experimentos. El rendimiento de FedPRC se evalúa frente a los métodos de referencia del estado del arte en cuatro conjuntos de datos heterogéneos, tal como se establece en la configuración experimental. Los resultados principales se resumen en la Tabla 1, que reporta la precisión y pérdida finales y promedio para cada método.

Tabla 1: Resultados obtenidos para los conjuntos de datos especificados, mostrando las métricas de precisión y pérdida. Los mejores resultados para cada conjunto de datos se marcan en negrita. Los segundos mejores resultados están subrayados.

Conjunto Datos	Método	Precisión Final	Pérdida Final	Precisión Prom.	Pérdida Prom.
CelebA-S	FedAvg	0.9040 ± 0.01	0.2828 ± 0.04	0.9052 ± 0.01	0.2680 ± 0.03
	FedProx	0.8982 ± 0.01	0.2523 ± 0.02	0.8902 ± 0.02	0.2731 ± 0.04
	FedNova	**0.9091 ± 0.00**	0.2481 ± 0.01	0.9057 ± 0.00	0.2634 ± 0.02
	FedDyn	0.8675 ± 0.03	0.3190 ± 0.07	0.8716 ± 0.02	0.3102 ± 0.05
	FedPRC	0.9080 ± 0.00	**0.2406 ± 0.01**	**0.9068 ± 0.00**	**0.2424 ± 0.01**
CelebA-A	FedAvg	0.6575 ± 0.12	0.7701 ± 0.39	0.6481 ± 0.09	0.7291 ± 0.22
	FedProx	0.6715 ± 0.09	0.6085 ± 0.11	**0.7000 ± 0.06**	**0.5893 ± 0.10**
	FedNova	0.6328 ± 0.11	0.7856 ± 0.33	0.6408 ± 0.10	0.7989 ± 0.32
	FedDyn	0.6863 ± 0.09	**0.5846 ± 0.10**	0.6513 ± 0.08	0.6750 ± 0.18
	FedPRC	**0.7127 ± 0.05**	0.5898 ± 0.08	0.6768 ± 0.07	0.6579 ± 0.15
CelebA-M	FedAvg	0.5951 ± 0.00	1.3467 ± 0.28	0.6258 ± 0.05	1.1750 ± 0.41
	FedProx	0.4994 ± 0.17	0.8190 ± 0.21	0.4940 ± 0.10	0.8229 ± 0.21
	FedNova	0.6664 ± 0.08	0.9561 ± 0.39	0.6299 ± 0.06	1.2453 ± 0.50
	FedDyn	0.4206 ± 0.03	1.0977 ± 0.24	0.4743 ± 0.11	1.0033 ± 0.32
	FedPRC	**0.7277 ± 0.11**	**0.5622 ± 0.16**	**0.7543 ± 0.09**	**0.5215 ± 0.15**
CIFAR-10	FedAvg	0.6841 ± 0.03	**0.9634 ± 0.09**	0.6607 ± 0.06	1.0231 ± 0.18
	FedProx	0.6080 ± 0.02	1.0350 ± 0.05	0.6056 ± 0.03	1.0586 ± 0.06
	FedNova	0.6732 ± 0.05	0.9963 ± 0.22	0.6767 ± 0.05	**0.9619 ± 0.17**
	FedDyn	0.0892 ± 0.02	7.8660 ± 2.55	0.0961 ± 0.02	6.7141 ± 1.57
	FedPRC	**0.6842 ± 0.04**	1.0058 ± 0.16	**0.6807 ± 0.03**	1.0486 ± 0.12

Rendimiento Comparativo (PI1). Analizando los resultados de la Tabla 1, observamos que FedPRC logra resultados competitivos o superiores en todos los escenarios evaluados. Específicamente, se posiciona de manera consistente como el método con mejor rendimiento o mantiene una distancia marginal con respecto al resultado superior (siendo la mayoría de las veces el segundo mejor método con un pequeño margen de diferencia frente al mejor rendimiento) en los cuatro entornos heterogéneos. Esta estabilidad es particularmente notable dado que los métodos de referencia exhiben fluctuaciones significativas de rendimiento dependiendo del conjunto de datos y de la naturaleza específica del sesgo de los datos. La ponderación basada en la relevancia de FedPRC actúa eficazmente como un regularizador global que garantiza que el límite de decisión permanezca robusto independientemente del cambio en la distribución subyacente.

Resiliencia ante Sesgos Mixtos (PI2). La suite de conjuntos de datos CelebA (CelebA-S, CelebA-M y CelebA-A) proporciona un banco de pruebas realista para entornos no-IID, ya que cada conjunto de celebridades incorpora una mezcla de sesgos de etiquetas, características y cantidad. Como se observa en la Tabla 1, FedPRC demuestra una resiliencia notable en estos escenarios complejos. En CelebA-M, nuestro enfoque proporciona un salto significativo en el rendimiento, superando a FedAvg por más de 13 puntos porcentuales en precisión y reduciendo eficazmente a la mitad la pérdida final en comparación con la mayoría de los métodos de referencia. También podemos verlo en el caso de CIFAR-10, donde bajo un sesgo de etiquetas severo, FedPRC es estable en situaciones donde una determinada clase puede no haber sido vista por ningún cliente. En este caso, nuestro método simplemente mantendría la frontera de decisión igual a la de la ronda anterior, dado que ningún cliente la ha actualizado.

Curiosamente, FedPRC demuestra una robustez significativa incluso al tratar con atributos altamente subjetivos como *Attractive* (Atractivo) en CelebA-A. A pesar de la ambigüedad conceptual inherente a esta etiqueta en comparación con atributos físicamente distintos, nuestra propuesta sigue siendo altamente competitiva. Esto subraya la capacidad de nuestra propuesta: aunque FedPRC promueve la alineación de conceptos visuales objetivos, retiene fuertes capacidades de generalización incluso cuando las características semánticamente plausibles son menos distintas, validando la versatilidad de nuestro esquema de ponderación basado en relevancia.

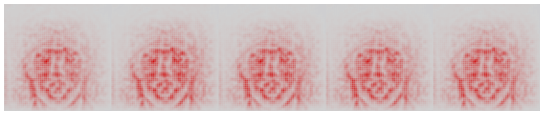
Robustez ante Desequilibrios Extremos (PI3). Para evaluar la robustez bajo un sesgo de etiquetas extremo, consideramos el conjunto de datos CIFAR-10 con una distribución de Dirichlet de $\alpha = 0.1$, cuyos resultados se muestran en la Tabla 1. En esta configuración, FedPRC alcanza una precisión al estado del arte (0.6842 ± 0.04). Una observación crítica es el colapso de FedDyn, que no logra converger y produce una precisión de aproximadamente el 9 %, rindiendo por debajo de un clasificador aleatorio. Mientras que FedDyn ha reportado éxito en entornos menos extremos ($\alpha \in \{0.3, 0.5\}$) [2], su fallo aquí resalta el peligro de una regularización dinámica agresiva en presencia de un sesgo de etiquetas severo. Más precisamente, observamos que la precisión se queda estancada alrededor

de 0.1 durante la mayoría de las rondas con perturbaciones sutiles, lo que denota el fracaso del aprendizaje. FedPRC, por el contrario, mantiene la estabilidad al anclar las representaciones al conjunto de datos del servidor D_S, evitando que el límite de decisión colapse.

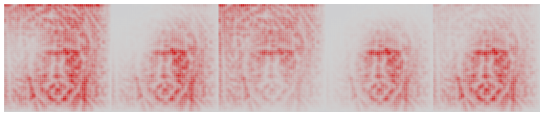
Alineación de Características e Interpretabilidad (PI4). Más allá de las métricas cuantitativas, investigamos si FedPRC conduce a una mejor alineación semántica del modelo global. Un subproducto considerable de nuestra agregación ponderada por relevancia es la producción de características más claras y más significativas desde el punto de vista semántico. Esto se confirma al analizar los mapas de calor de CRP en la Fig. 1 para el atributo "Smiling" (Sonriente) en CelebA-S. Para que una muestra se clasifique correctamente como "Sonriente", la relevancia debería concentrarse alrededor de la boca y la parte inferior del rostro del sujeto. Como se muestra en la figura, FedPRC produce mapas de calor que están estrictamente alineados con estos conceptos intrínsecos de la tarea. En cambio, FedNova, que es el método de referencia más cercano, produce mapas de calor que se enfocan en el fondo o el cabello. Esto sugiere que la guía semántica proporcionada por FedPRC mejora el enfoque en las características semánticamente plausibles. Esta propiedad es crucial para el despliegue del AF en dominios críticos para la seguridad como la salud, donde la interpretabilidad y fidelidad del modelo son primordiales.



(a) Imagen original de la partición de prueba de CelebA.



(b) Características de FedPRC en CelebA-S



(c) Características de FedNova en CelebA-S

Figura 1: Mapas de calor de las 5 características más relevantes para la clasificación correcta de una muestra dada, extraídos con CRP. FedPRC ofrece un mapa más útil, mostrando una alineación adecuada de las características para permitir la generalización.

5.3. Limitaciones y Consideraciones Prácticas

Aunque FedPRC proporciona un rendimiento empírico robusto, depende de que el servidor disponga de un pequeño conjunto de validación etiquetado D_S.

Esto presenta ciertas limitaciones: (1) Consideraciones de privacidad: La presencia de D_S puede desviarse de los modelos puros de AF. Sin embargo, esto es altamente factible en el AF cross-silo (por ejemplo, hospitales, bancos, etc.), donde a menudo se dispone de datos de validación. Además, D_S nunca sale del servidor, preservando la privacidad del cliente. (2) Sesgo de selección de datos: Las muestras elegidas para D_S influyen en qué conceptos se agregan. (3) Afirmaciones de alineación semántica: Nuestra evaluación de la alineación semántica (PI4) es principalmente cualitativa, ya que el cálculo de métricas de explicabilidad cuantitativas (por ejemplo, puntuaciones de inserción/eliminación o pointing game) en todos los clientes y datos introduce un cuello de botella computacional masivo.

A pesar de estas advertencias, los conjuntos de datos de validación son altamente prácticos y se utilizan comúnmente en entornos reales de AF para depuración, validación o cumplimiento. FedPRC aprovecha este recurso existente para mejorar la alineación del modelo global.

6. Conclusiones

Este trabajo presenta FedPRC, un mecanismo de optimización novedoso que utiliza la CRP para resolver la deriva del modelo bajo heterogeneidad estadística de datos en el AF. Al anclar los pesos de los clientes a un conjunto de validación del servidor D_S y calcular atribuciones a nivel de concepto mediante CRP, FedPRC agrega la cabeza del clasificador potenciando selectivamente los conceptos universalmente relevantes y suprimiendo los artefactos específicos de los clientes.

FedPRC logra un rendimiento competitivo o superior en todos los conjuntos de datos evaluados, ofreciendo ganancias de precisión de hasta el 13 % en la suite CelebA y manteniendo la estabilidad (precisión del 68.4 % en CIFAR-10, $\alpha = 0.1$) donde las regularizaciones dinámicas del estado del arte colapsan. Además, nuestro enfoque alinea de forma natural las características globales con los conceptos intrínsecos de la tarea (por ejemplo, la parte inferior del rostro para la clasificación de sonrisas), demostrando que los mecanismos de explicabilidad post-hoc pueden actuar como ingredientes activos durante el propio proceso de aprendizaje. Las direcciones futuras incluyen la exploración de métodos alternativos de atribución XAI (como SHAP o pesos de atención) y la vinculación de FedPRC con explicaciones causales basadas en conceptos [7,13,6].

Agradecimientos Esta investigación es parte del Proyecto TSI-100927-2023-1, financiado por Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-Next Generation EU a través del Ministerio para la Transformación Digital y de la Función Pública y está parcialmente financiada por el Ministerio de Ciencia y Tecnología de España en el marco del proyecto PID2023-150070NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE, y por el proyecto PIN-FLEX del Centro Mixto UGR-MADOC.

Declaración de Intereses. Los autores declaran no tener conflictos de intereses que sean relevantes para el contenido de este artículo.

Referencias

1. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance) (May 2016)
2. Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: International Conference on Learning Representations (2021)
3. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence **5**(9), 1006–1019 (2023)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
5. Barona López, L.I., Borja Saltos, T.: Heterogeneity challenges of federated learning for future wireless communication networks. Journal of Sensor and Actuator Networks **14**(2), 37 (2025)
6. Bjøru, A.R., Lysnæs-Larsen, J., Jørgensen, O., Strümke, I., Langseth, H.: A framework for causal concept-based model explanations. arXiv preprint arXiv:2512.02735 (2025)
7. Carloni, G., Berti, A., Colantonio, S.: The role of causality in explainable artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **15**(2), e70015 (2025)
8. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International conference on machine learning. pp. 2089–2099. PMLR (2021)
9. of European Union, C.: Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Herrera, F., Jiménez-López, D., Argente-Garrido, A., Rodríguez-Barroso, N., Zuheros, C., Aguilera-Martos, I., Bello, B., García-Márquez, M., Luzón, M.: Flex: Flexible federated learning framework. Information Fusion **117**, 102792 (2025)
12. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and trends® in machine learning **14**(1–2), 1–210 (2021)
13. Karimi, A.H.: Position: Explainable ai is causality in disguise. arXiv preprint arXiv:2603.28597 (2026)
14. Kirichenko, P., Izmailov, P., Wilson, A.G.: Last layer re-training is sufficient for robustness to spurious correlations. In: The Eleventh International Conference on Learning Representations (2023)
15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
16. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)

17. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
18. Luzón, M.V., Rodríguez-Barroso, N., Argente-Garrido, A., Jiménez-López, D., Moyano, J.M., Ser, J.D., Ding, W., Herrera, F.: A tutorial on federated learning from theory to practice: Foundations, software frameworks, exemplary use cases, and selected trends. IEEE/CAA Journal of Automatica Sinica **11**(JAS-2023-0937), 824 (2024)
19. Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J.C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J.B., Rome, A., Shi, A., Oak, S.: Artificial intelligence index report 2025 (2025)
20. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017)
21. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce, Gaithersburg, MD (2023), nIST Special Publication AI 100-1
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
23. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems **33**, 7611–7623 (2020)