Vers une protection de la vie privée équitable pour les services de santé

Auteurs

Adda Akram BENDOUKHA

Nesrine KAANICHE

Renaud SIRDEY

Aymen BOUDGUIGA

Partenaires

Soutenu par

Projet EQUIHID

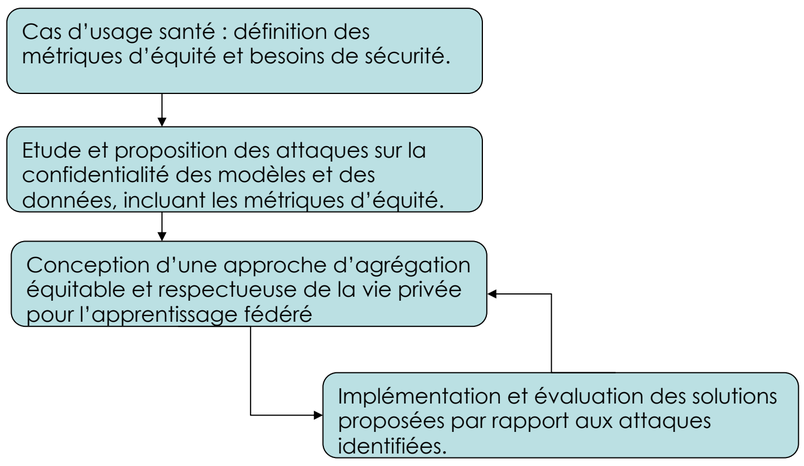
MOTIVATION ET OBJECTIFS

- Des systèmes de santé intelligents reposant sur la collecte massive de données sensibles
  - ➢ Plusieurs lois et régulations : RGPD, AI Act, e-privacy,
  - ➢ Absence de législation sur l'équité
- Des applications basées sur l'IA développées avec un souci limité de la vie privée des utilisateurs
  - ➢ Reconstruction des données, inférence d'informations sensibles, reconstitution des profils des patients.
- Difficile d'utiliser de larges bases de données, équilibrées et diverses, nécessaires pour garantir de meilleurs résultats de classification
- Des algorithmes d'apprentissage distribués développés sans considération majeure des problèmes d'équité
  - ➢ Des données d'entrainement conduisant à des comportements discriminatoires et/ou amplifiant des biais sociétaux.
  - ➢ Des algorithmes augmentant des risques d'identification des classes minoritaires de 20%.
  - ➢ Divulguer les métriques d'équité permettant d'améliorer le taux de succès des attaques par inférence

⇒ **Assurer un compromis entre équité, respect de la vie privée et utilité.**

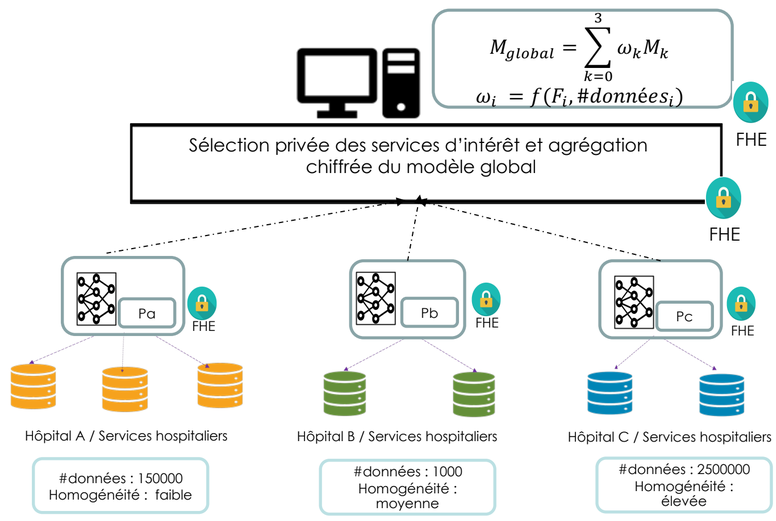
MÉTHODOLOGIE

- Proposer une nouvelle approche d'**apprentissage fédéré**, répondant aux métriques d'**équité** et assurant une **agrégation** respectueuse de la vie **privée**.
- Concevoir de nouvelles attaques sur la confidentialité des modèles et des données et combinant des métriques d'équité dans un contexte d'apprentissage distribué.
- Evaluer l'efficacité de l'approche proposée par rapport aux attaques implémentées.



NOTRE APPROCHE

CHIFFREMENT HOMOMORPHE POUR UN APPRENTISSAGE FÉDÉRÉ EQUITABLE ET SÉCURISÉ



Etat de l'art: Les fonctions d'agrégation des protocoles d'apprentissage fédéré ne prennent pas en compte le niveau d'équité des modèles générés au niveau des clients.

Problème 1: Un modèle-client *peu équitable* contribuera à part égal à la génération du modèle global qu'un modèle équitable, amplifiant ainsi les biais.

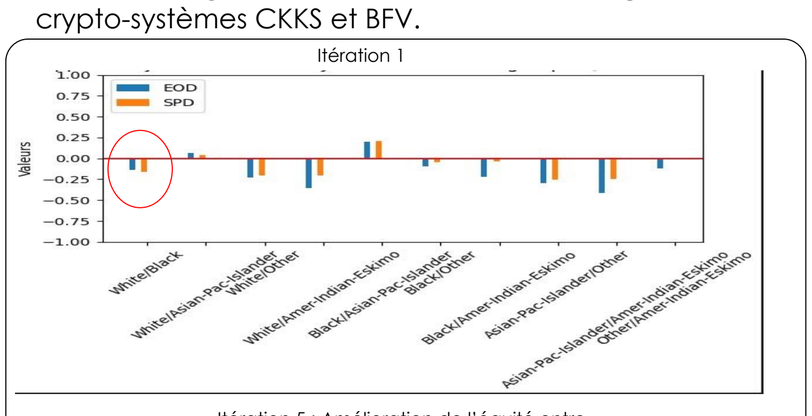
⇛ **Solution:** Ajouter une **pondération proportionnelle à un score d'équité**, calculé au niveau chaque modèle-client, lors de l'agrégation du modèle global.

Problème 2: La révélation d'une métrique d'équité du modèle constitue un risque de confidentialité.

⇛**Solution:** Utiliser le chiffrement homomorphe, *multi-clés,* (**CKKS & BFV**) [2] pour protéger à la fois les mises à jour du modèles ainsi que les métriques d'équité individuelles et de groupe.

RÉSULTATS

- Une première approche prometteuse pour un modèle d'attaque: Serveur et Clients honnête(s) mais curieux
  - ➢ Une amélioration de l'équité de groupe (EOD et SPD) du modèle agrégé par rapport à une agrégation dont la pondération n'est pas liée aux scores d'équité.
  - ➢ Les calculs homomorphes : (évaluation de la fonction f + agrégation) permettront d'avoir de résultats prometteurs grâce à la méthode du **Batching** des crypto-systèmes CKKS et BFV.



Itération 5 : Amélioration de l'équité entre les groupes Black/White pour l'attribut

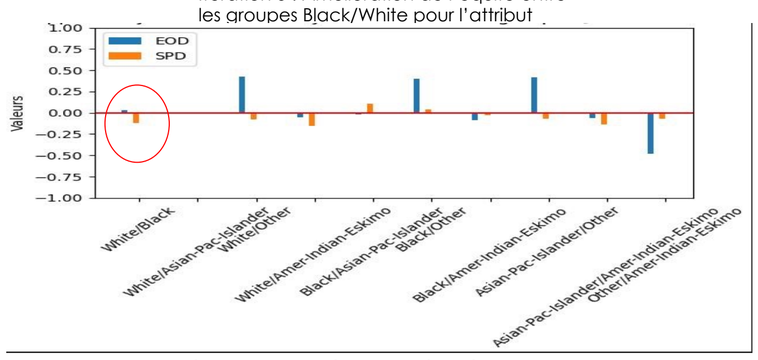


Illustration de l'évolution des métriques d'équité durant le processus de FL avec une fonction d'agrégation respectueuse de l'équité. La cible est de minimiser les écarts entre les groupes **White/Black** dans la dataset Adult Census Income

$$SPD(M) = P(\hat{Y} = 1 \mid S = s_0) - P(\hat{Y} = 1 \mid S = s_1)$$
$$EOD(M) = P(\hat{Y} = 1 \mid Y = 1, S = s_0) - P(\hat{Y} = 1 \mid Y = 1, S = s_1)$$

- SPD : Statistical Parity Difference (Valeur idéale 0)
- EOD : Equal Opportunity Difference (Valeur idéale 0)

RÉFÉRENCES

1. A-A Bendoukha, A. Boudguiga, R. Sirdey, *Homomorphic Sortition - Single Secret Leader Election for PoS Blockchains* IACR Cryptol. ePrint Arch. 2023: 113 (2023),
2. L. Freitas de Souza, A. Tonkikh, **A-A Bendoukha**, S. TucciPiergiovanni, R. Sirdey, Oana Stan, P. Kuznetsov, *Revisiting Stream-Cipher-Based Homomorphic Transciphering in the TFHE Era*. FPS 2021: 19-33
3. A -A Bendoukha O. Stan, R. Sirdey, N. Quero, and L. Freitas, *Practical homomorphic evaluation of block-cipher-based hash functions with applications* FPS 2023.

Contact adda-akram.bendoukha@telecom-sudparis.eu

@equihid