Apprentissage Fédéré

Implémentation et analyse sur le jeu de données MNIST

Par

Clément COLIN

Enzo ROCAMORA

Thomas CHOUBRAC

Avril 2025

Sommaire

Introduction

L'apprentissage fédéré répond à plusieurs enjeux. D'une part, il respecte la confidentialité des données, sujet sensible depuis l'entrée en vigueur du RGPD. D'autre part, il permet de tirer parti de données distribuées géographiquement sans avoir à les centraliser, ce qui peut s'avérer impossible pour des raisons techniques, légales ou d'évolution constante des données.

Notre objectif a été d'implémenter et d'analyser différentes variantes d'apprentissage fédéré sur le jeu de données MNIST, afin de comprendre les impacts de divers paramètres sur les performances des modèles.

Fondements théoriques de l'apprentissage fédéré

Principe général

L'apprentissage fédéré se déroule généralement en trois étapes principales :

1. Un serveur central initialise un modèle global et le distribue aux clients
2. Chaque client entraîne le modèle sur ses données locales
3. Les mises à jour des modèles locaux sont agrégées par le serveur central

Mathématiquement, l'objectif est de résoudre un problème d'optimisation de la forme :

$$\min_{w \in \mathbb{R}^d} f(w)$$

où $f(w)$ est décomposable sous forme d'une somme finie :

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} f_i(w)$$

Dans le contexte du machine learning, la fonction f_i représente la fonction de perte pour le point de données i :

$$f_i(w) = \text{loss}(x_i, y_i; w)$$

Lorsque les données sont réparties entre K clients, nous pouvons réécrire l'objectif comme :

$$f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w)$$

où $F_k(w) = \frac{1}{n_k} \sum_{i \in P_k} f_i(w)$ est la fonction objectif locale du client k, P_k est l'ensemble des indices des points de données du client k, $n_k = |P_k|$ est le nombre de données du client k, et $n = \sum_{k=1}^{K} n_k$ est le nombre total de données.

Jeu de données MNIST

Pour notre étude, nous avons utilisé le jeu de données MNIST, qui contient des images en niveaux de gris de chiffres manuscrits (de 0 à 9). Voici ses caractéristiques principales :

- 37 800 images d'entraînement
- 4 200 images de test
- Dimensions des images : 28×28 pixels (en niveaux de gris)

La distribution des classes dans l'ensemble d'entraînement est relativement équilibrée, avec environ 10% des images pour chaque chiffre. Cette répartition équilibrée nous a permis d'expérimenter facilement différentes stratégies de distribution entre clients.

Pour charger et prétraiter ces données, nous avons implémenté le module `fl_dataquest.py` qui convertit les images en tenseurs normalisés et crée des pipelines TensorFlow efficaces :

```
def get_data(img_path = '../Mnist/trainingSet/trainingSet/', verbose = 0):
    # Chargement et prétraitement des images MNIST
    image_paths = list(list_images(img_path))
    il, ll = load_and_preprocess(image_paths, verbose=10000)

    # Binarisation des labels (one-hot encoding)
    lb = skl.preprocessing.LabelBinarizer()
    ll = lb.fit_transform(ll)

    # Division en ensembles d'entraînement et de test
    X_train, X_test, y_train, y_test = train_test_split(il, ll,
                                                        test_size=0.1,
                                                        random_state=19)

    return X_train, X_test, y_train, y_test, X_train[0].shape
```

Tentative d'application sur d'autres jeux de données

Nous avons également exploré l'utilisation d'autres jeux de données, notamment le dataset Auto-MPG (Session 2). Mais nous avons rencontré plusieurs obstacles. La taille réduite du jeu de données rendait difficile une distribution significative entre plusieurs clients. De plus, les caractéristiques numériques continues présentaient des distributions très hétérogènes, ce qui entraînait des problèmes de convergence dans le cadre de l'apprentissage fédéré. Nos tests préliminaires ont montré des écarts de performance considérables entre les différents clients, selon la distribution des données.

Architecture du modèle et méthodologie expérimentale

Prérequis et installation

Pour exécuter notre implémentation d'apprentissage fédéré, vous aurez besoin d'installer les bibliothèques Python suivantes.

```
pip install tensorflow numpy matplotlib scikit-learn opencv-python pandas
seaborn psutil
```

Structure du projet

Notre projet est organisé en plusieurs modules Python :

- `fl_model.py` : Définition du modèle neuronal
- `fl_types.py` : Implémentation de l'apprentissage fédéré horizontal
- `fl_dataquest.py` : Chargement et prétraitement des données MNIST
- `data_partition.py` : Fonctions pour la répartition des données entre clients
- `aggregation.py` : Algorithmes d'agrégation (FedAvg, FedSGD, FedProx)
- `utils.py` : Fonctions utilitaires
- `run_all.py` : Script principal pour exécuter toutes les expériences

Exécution des expériences

Pour lancer l'ensemble des expériences, exécutez simplement le script principal :

```
python run_all.py
```

Ce script va :

1. Charger les données MNIST
2. Créer un modèle centralisé de référence
3. Exécuter toutes les configurations d'apprentissage fédéré définies
4. Évaluer les performances de chaque configuration
5. Générer des visualisations et sauvegarder les résultats

L'exécution complète peut prendre plusieurs heures selon votre matériel, car de nombreuses configurations sont testées (nombre de clients, de rounds, d'époques locales, différents algorithmes, etc.).

Fichiers générés

1. **Résultats CSV** :
   - `federated_learning_results.csv` : Résumé des performances finales de chaque configuration
   - `federated_learning_detailed_results.csv` : Résultats détaillés round par round
2. **Visualisations** :
   - Dossier `plots/` : Graphiques comparatifs des différentes configurations
3. **Visualisations de progression** :
   - Dossier `progression_plots/` : Visualisations PCA de la trajectoire des paramètres
4. **Historique des paramètres** :
   - Dossier `parameter_history/` : Fichiers pickle contenant l'historique des paramètres pour chaque configuration

Modèle neuronal

Pour notre étude, nous avons utilisé un réseau de neurones multicouche (MLP) implémenté dans la classe `MyModel` du module `fl_model.py` :

```
class MyModel():
    def __init__(self, input_shape, nbclasses):
        model = Sequential()
        model.add(Input(input_shape))
        model.add(Flatten())
        model.add(Dense(128))
        model.add(Activation("relu"))
        model.add(Dense(64))
        model.add(Activation("relu"))
        model.add(Dense(nbclasses))
        model.add(Activation("softmax"))

        self.model = model
        self.input_shape = input_shape
        self.classes = nbclasses

        self.loss_fn = 'categorical_crossentropy'
           self.model.compile(optimizer="SGD", loss=self.loss_fn, metrics=
["accuracy"])
```

Cette architecture relativement simple est suffisante pour obtenir de bons résultats sur MNIST tout en restant légère en termes de calcul, ce qui est important dans un contexte d'apprentissage fédéré

où les ressources des clients peuvent être limitées.

Mécanisme d'apprentissage fédéré

Le cœur de notre implémentation repose sur la fonction `horizontal_federated_learning` du module `fl_types.py`, qui coordonne l'entraînement entre le serveur central et les clients :

```
def horizontal_federated_learning(edges, central_model, input_shape, num_classes,
                                  edge_epochs, test_data, aggregation_fn, verbose=0):
    '''Apprentissage fédéré horizontal (HFL)'''
    central_weights = central_model.get_weights()
    scaled_local_weight_list = []

    # Pour chaque client
    edges_names = list(edges.keys())
    random.shuffle(edges_names)

    for client_name in edges_names:
        # Obtenir les données du client
        client_data = edges[client_name]

        # Créer et configurer le modèle local
        local_model = fl_model.MyModel(input_shape, nbclasses=num_classes)
        local_model.set_weights(central_weights)

        # Entraîner le modèle local
                local_model.fit_it(trains=client_data, epochs=edge_epochs, tests=test_data, verbose=verbose)

        # Calculer le facteur d'échelle
        scaling_factor = _weight_scaling_factor(edges, client_name)
          scaled_weights = _scale_model_weights(local_model.get_weights(), scaling_factor)
        scaled_local_weight_list.append(scaled_weights)

        # Nettoyer la session
        K.clear_session()

    # Agréger les poids et mettre à jour le modèle central
    updated_weights = aggregation_fn(scaled_local_weight_list)
    central_model.set_weights(updated_weights)

    return central_model
```

Cette fonction implémente les étapes clés de l'apprentissage fédéré horizontal :

1. Distribution du modèle global à tous les clients
2. Entraînement local sur chaque client pendant un nombre spécifié d'époques
3. Calcul d'un facteur d'échelle pour chaque client proportionnel à la quantité de données dont il dispose
4. Agrégation des poids mis à l'échelle via une fonction d'agrégation spécifiée
5. Mise à jour du modèle central avec les poids agrégés

Distribution des données entre clients

Un aspect essentiel de notre étude était d'expérimenter avec différentes stratégies de distribution des données entre clients. Nous avons implémenté trois types de distribution :

1. **Distribution IID** (Independent and Identically Distributed) : les données sont réparties aléatoirement entre les clients, garantissant une distribution similaire des classes pour chaque client.

```
def iid_partition(X, y, num_clients, batch_size=32):
      '''Distribution IID: chaque client a des données indépendantes et
identiquement distribuées'''
    data = list(zip(X, y))
    random.shuffle(data)

    size = len(data)//num_clients
    shards = [data[i:i + size] for i in range(0, size*num_clients, size)]

    client_data = {}
    for i in range(num_clients):
        client_name = f'edge_{i}'

        client_X, client_y = zip(*shards[i])
        client_X, client_y = list(client_X), list(client_y)

        dataset = tf.data.Dataset.from_tensor_slices((client_X, client_y))
        dataset = dataset.shuffle(len(client_y))
        dataset = dataset.batch(batch_size)

        client_data[client_name] = dataset

    return client_data
```

2. **Distribution non-IID** : chaque client reçoit un sous-ensemble biaisé des classes, créant une hétérogénéité dans la distribution des données.

```
def non_iid_partition(X, y, num_clients, classes_per_client=2, batch_size=32):
    '''Distribution non-IID: chaque client a un sous-ensemble déséquilibré des classes'''
    labels = np.argmax(y, axis=1)
    sorted_indices = np.argsort(labels)

    X_sorted = [X[i] for i in sorted_indices]
    y_sorted = [y[i] for i in sorted_indices]

    # Compter combien de classes nous avons
    num_classes = y[0].shape[0]
    samples_per_class = len(X) // num_classes

    # Répartir les données entre les clients
    client_data = {}
    for i in range(num_clients):
        client_name = f'edge_{i}'
        client_X, client_y = [], []

        # Pour chaque client, sélectionner classes_per_client classes
        selected_classes = np.random.choice(range(num_classes), classes_per_client, replace=False)

        # ... [code pour allouer les données selon les classes sélectionnées] ...

        # Créer le dataset pour ce client
        dataset = tf.data.Dataset.from_tensor_slices((client_X, client_y))
        dataset = dataset.shuffle(len(client_y))
        dataset = dataset.batch(batch_size)

        client_data[client_name] = dataset

    return client_data
```

3. **Distribution non-IID extrême** : chaque client se spécialise presque exclusivement sur une ou deux classes, avec très peu d'exemples des autres classes.

Ces différentes distributions nous permettent d'étudier la robustesse des algorithmes d'apprentissage fédéré face à l'hétérogénéité des données, un défi majeur dans les applications réelles.

Algorithmes d'agrégation

Nous avons implémenté et comparé trois algorithmes d'agrégation principaux :

1. **FedAvg** (Federated Averaging) : l'algorithme standard qui calcule une moyenne pondérée des poids des modèles locaux.

```
def fedavg(scaled_weight_list, central_weights=None, config_name=None, round_num=None):
    avg_weights = list()
    for grad_list_tuple in zip(*scaled_weight_list):
        layer_mean = tf.math.reduce_sum(grad_list_tuple, axis=0)
        avg_weights.append(layer_mean)

    return avg_weights
```

2. **FedSGD** (Federated Stochastic Gradient Descent) : agrégation basée sur les gradients plutôt que sur les poids eux-mêmes.

```
def fedsgd(model, client_grads, learning_rate=0.01, central_weights=None, config_name=None, round_num=None):
    current_weights = model.get_weights()
    updated_weights = []
    for i in range(len(current_weights)):
            updated_weights.append(current_weights[i] - learning_rate * avg_grads[i])

    return updated_weights
```

3. **FedProx** (Federated Proximal) : une extension de FedAvg qui ajoute un terme de régularisation proximal pour limiter la divergence entre les modèles.

```
def fedprox(scaled_weight_list, global_weights, mu=0.01, config_name=None, round_num=None):
    avg_weights = []
    for grad_list_tuple in zip(*scaled_weight_list):
        layer_mean = tf.math.reduce_sum(grad_list_tuple, axis=0)
        avg_weights.append(layer_mean)
    for i in range(len(avg_weights)):
        proximal_term = mu * (avg_weights[i] - global_weights[i])
        avg_weights[i] = avg_weights[i] - proximal_term

    return avg_weights
```

Paramètres étudiés et expérimentations

Dans nos expériences, nous avons fait varier plusieurs paramètres clés :

1. **Nombre de rounds fédérés** (3, 5, 10) : combien de fois les clients et le serveur échangent des mises à jour.
2. **Nombre de clients** (5, 10, 20) : combien d'entités participent à l'apprentissage fédéré.
3. **Distribution des données** (IID, non-IID, non-IID extrême) : comment les données sont réparties entre les clients.
4. **Nombre d'époques locales** (1, 3, 5) : combien d'époques d'entraînement chaque client effectue localement.
5. **Algorithmes d'agrégation** (FedAvg, FedSGD avec différents taux d'apprentissage, FedProx avec différentes valeurs du paramètre μ).

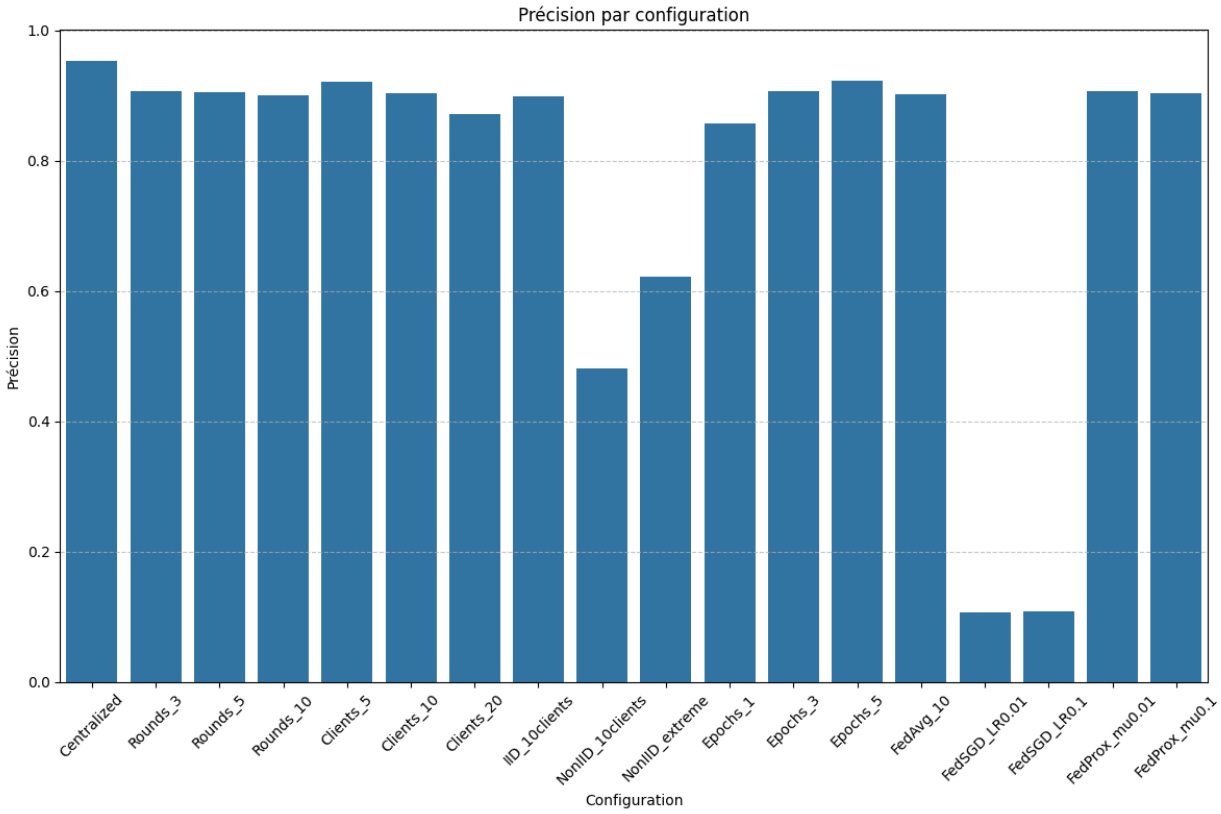
Ces expériences ont été exécutées via le script `run_all.py`, qui implémente une boucle complète d'expérimentation :

```
def run_all_experiments():
    # ... [initialisation] ...

    configurations = [
           {'name': 'Rounds_3', 'num_clients': 10, 'distribution': 'iid',
'algo': 'fedavg', 'epochs': 3, 'rounds': 3},
           {'name': 'Rounds_5', 'num_clients': 10, 'distribution': 'iid',
'algo': 'fedavg', 'epochs': 3, 'rounds': 5},
        # ... [autres configurations] ...
    ]

    for config in configurations:
        # ... [répartition des données selon la configuration] ...

        for round_num in range(num_rounds):
            federated_model = fl_types.horizontal_federated_learning(
                edges=edges,
                central_model=federated_model,
                input_shape=input_shape,
                num_classes=10,
                edge_epochs=config['epochs'],
                test_data=test_dataset,
                aggregation_fn=agg_fn,
                verbose=0
            )

             round_loss, round_acc = federated_model.evaluate(test_dataset,
verbose=0)
            # ... [enregistrement des résultats] ...
```

Résultats et analyse

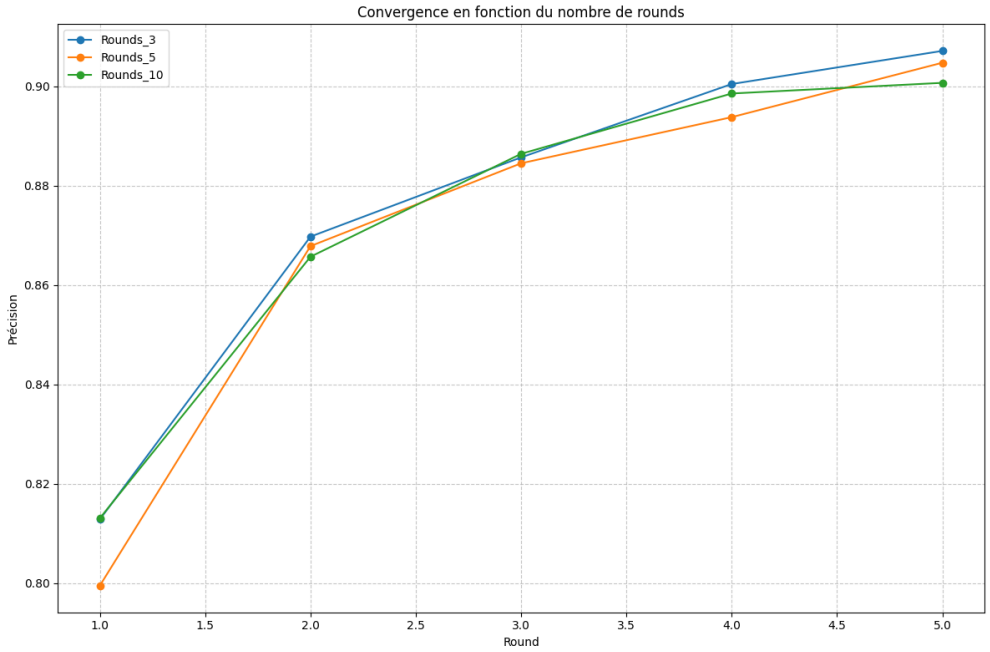
Comparaison avec l'entraînement centralisé

Notre modèle de référence, entraîné de manière centralisée avec accès à toutes les données, atteint une précision de 95% après 10 époques. Aucune des configurations fédérées n'atteint tout à fait cette performance, ce qui était attendu étant donné les contraintes de l'apprentissage distribué. Cependant, les meilleures configurations fédérées parviennent à des performances respectables, jusqu'à 92% dans certains cas.



Influence du nombre de rounds fédérés

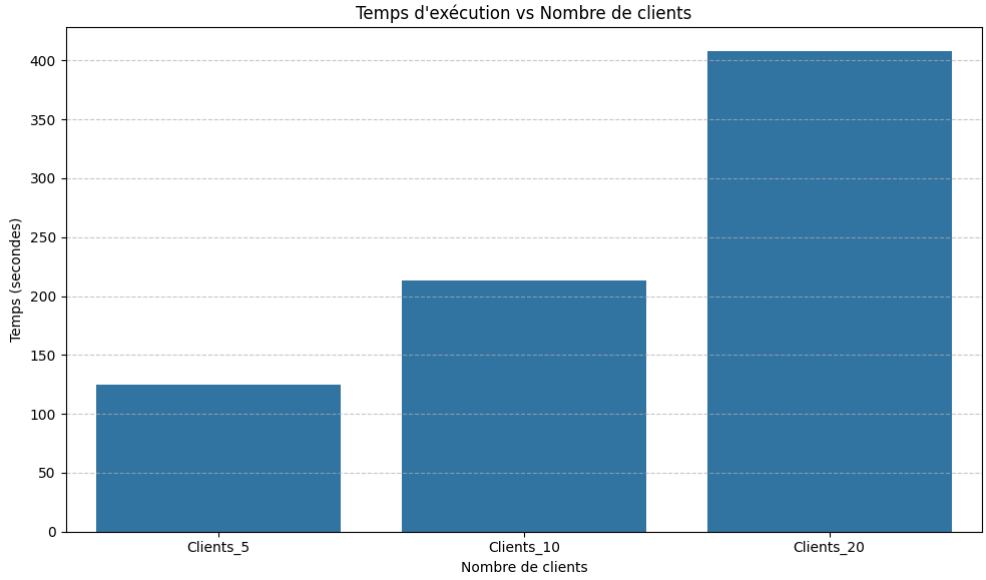
Nos expériences montrent que la précision augmente avec le nombre de rounds, mais avec des rendements décroissants. Après 5 rounds, les configurations avec 3, 5 et 10 rounds atteignent respectivement 90,7%, 90,5% et 90,1% de précision. Cette similitude suggère qu'au-delà de 5 rounds, le gain de performance devient marginal pour cette tâche.



La convergence suit une courbe typique d'apprentissage : rapide au début, puis qui ralentit progressivement. Cette observation est importante pour optimiser les ressources dans un déploiement réel.

Impact du nombre de clients

Nous avons observé une relation inverse entre le nombre de clients et la précision du modèle final. Avec 5 clients, la précision atteint 92%, tandis qu'avec 10 et 20 clients, elle baisse respectivement à 90% et 87%.

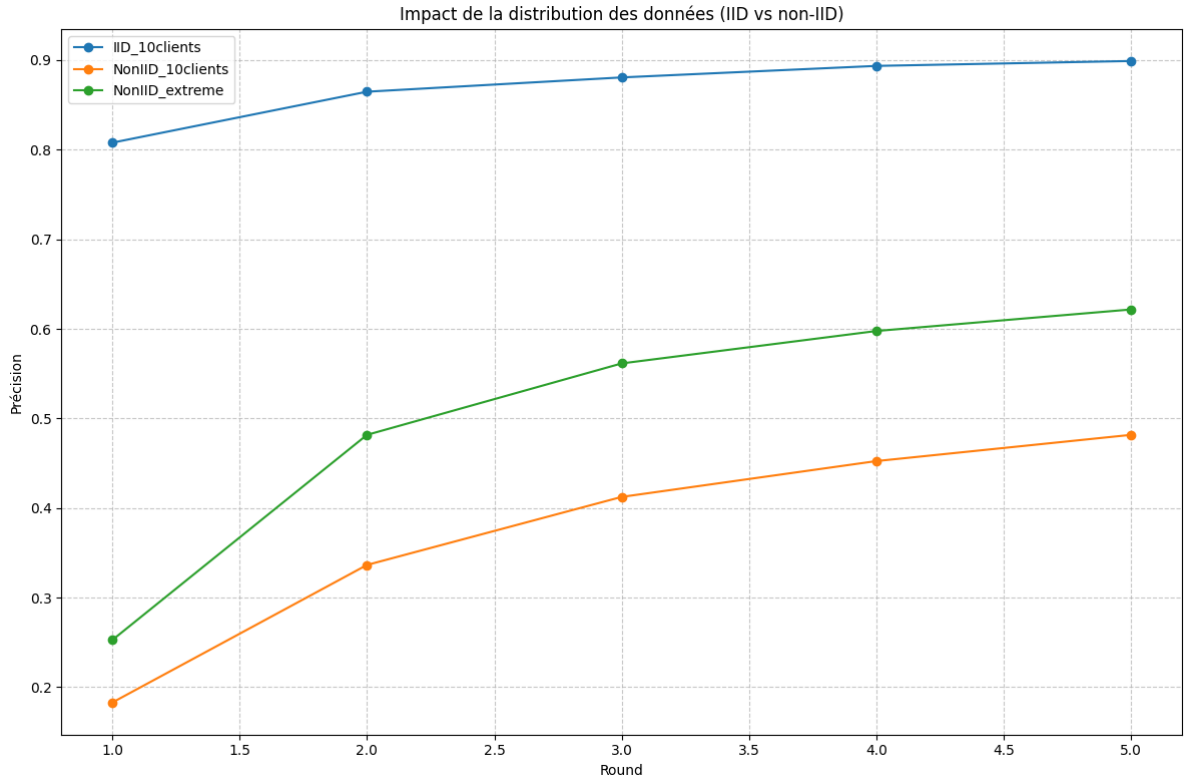

Temps d'exécution vs Nombre de clients
Temps (secondes)
0
50
100
150
200
250
300
350
400
Clients_5
Clients_10
Clients_20
Nombre de clients

Cette dégradation peut s'expliquer par plusieurs facteurs :

1. Moins de données par client, limitant la capacité d'apprentissage individuelle
2. Plus grande diversité de mises à jour, créant potentiellement des interférences lors de l'agrégation
3. Augmentation quasi-linéaire du temps de calcul (125s pour 5 clients, 213s pour 10 clients, 408s pour 20 clients)

Influence de la distribution des données

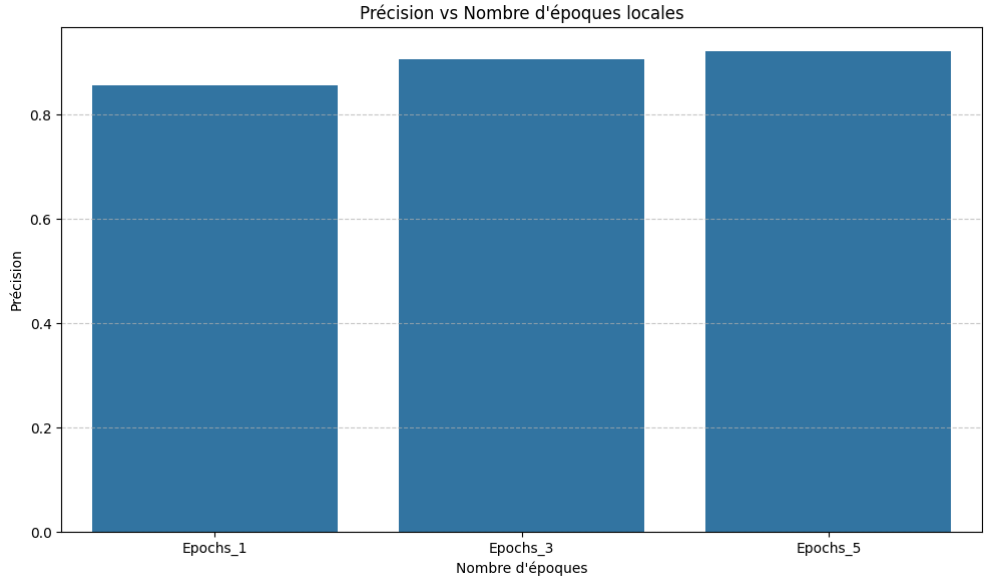
La distribution des données entre clients s'est avérée être le facteur le plus critique. En configuration IID, le modèle atteint 90% de précision. En revanche, en configuration non-IID, la performance chute drastiquement à 48,2%.

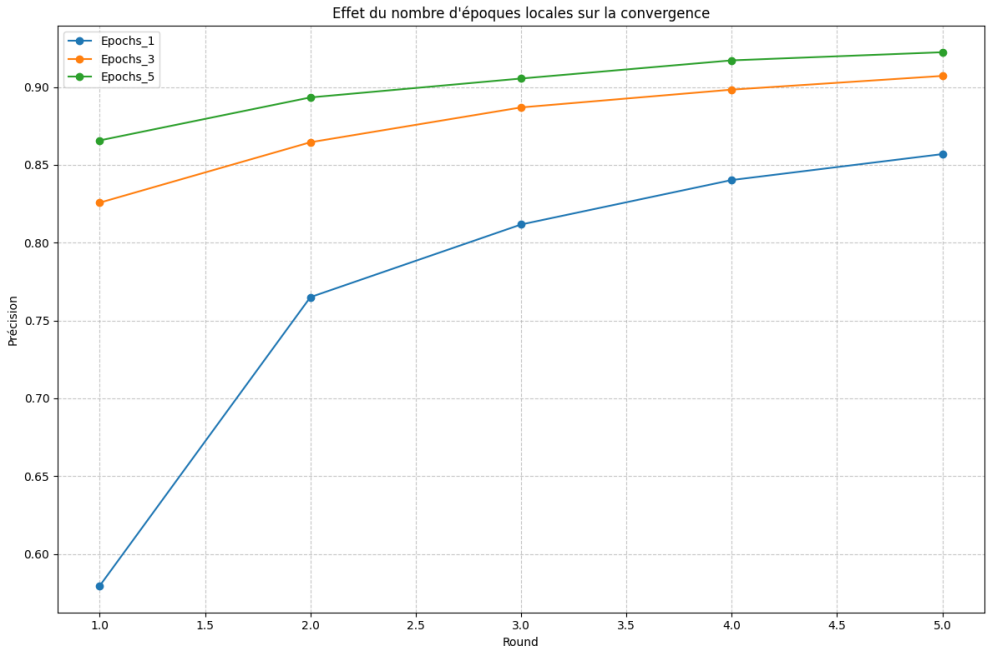


Fait intéressant, dans le scénario non-IID extrême, la précision remonte à 62%. Cette observation contre-intuitive pourrait s'expliquer par le fait que chaque client devient "expert" dans la reconnaissance de classes spécifiques, compensant partiellement les défis de l'hétérogénéité.

Effet du nombre d'époques locales

Avec une seule époque locale, la précision finale atteint 85,7%. Ce chiffre augmente à 90,7% avec 3 époques et à 92,2% avec 5 époques. Cette progression montre qu'un plus grand nombre d'époques permet à chaque modèle local de mieux apprendre à partir de ses données.

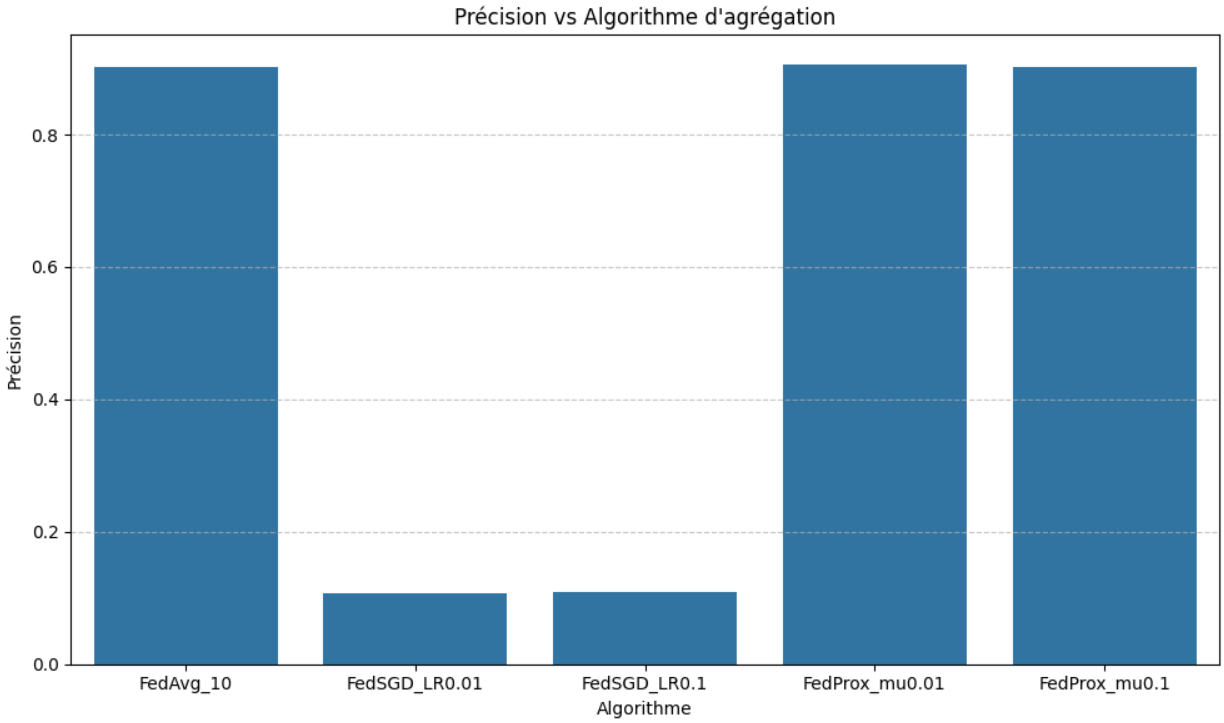




Toutefois, ce gain se fait au prix d'un temps de calcul plus élevé : 189s pour 1 époque, 249s pour 3 époques, et 295s pour 5 époques. Il existe également un risque qu'un trop grand nombre d'époques conduise à une sur-spécialisation des modèles sur leurs données locales.

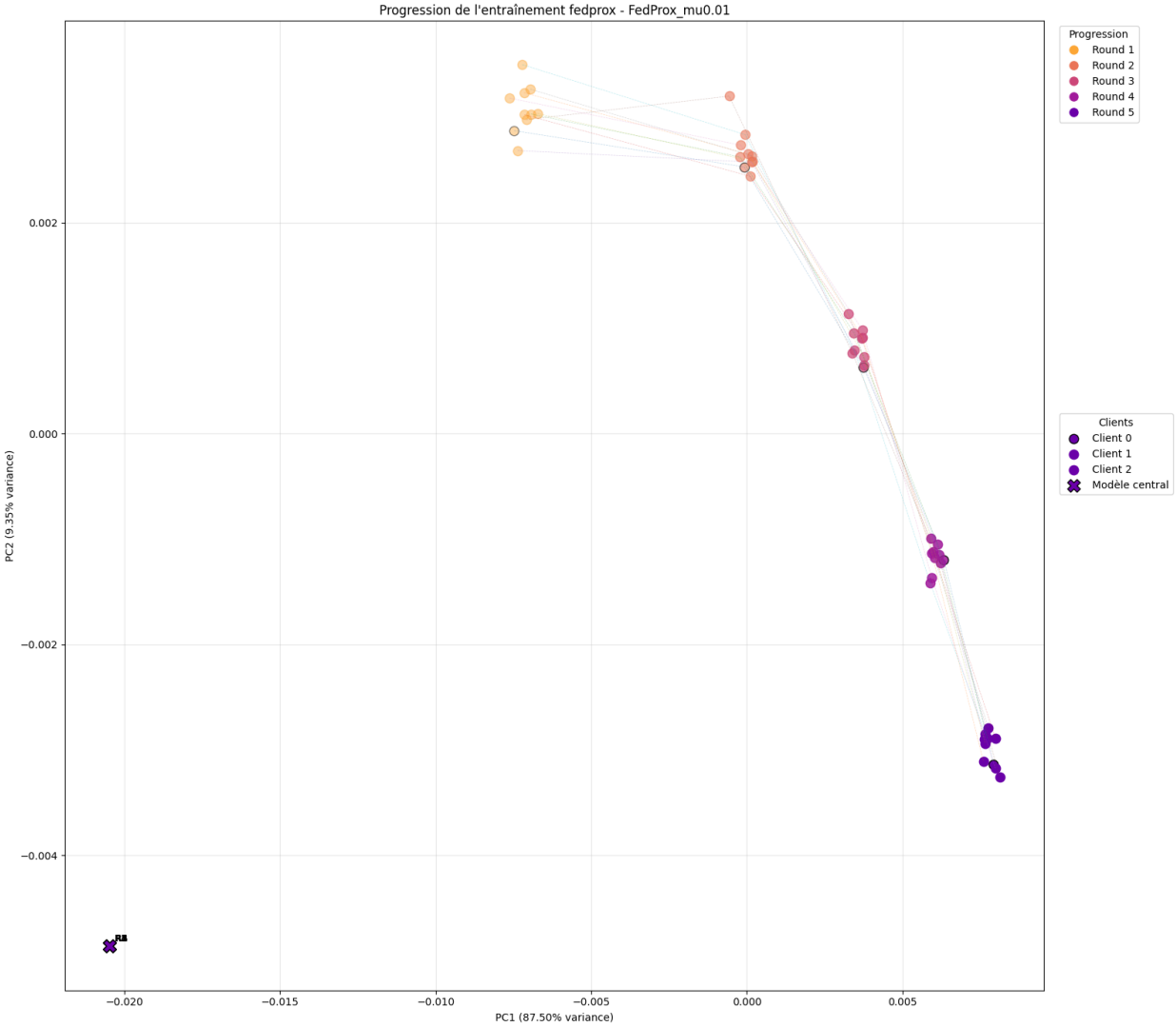
Comparaison des algorithmes d'agrégation

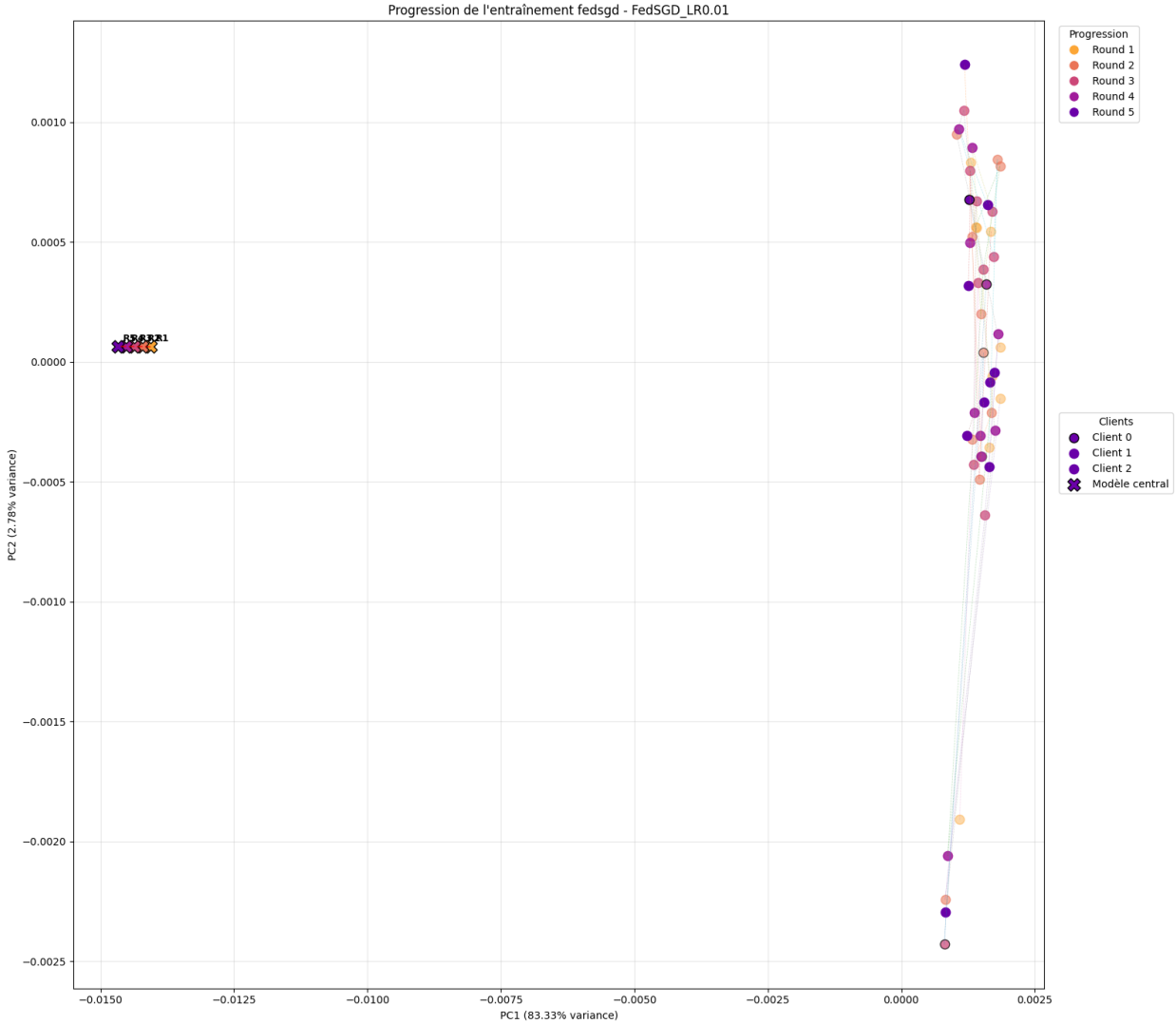
FedAvg, avec une précision de 90%, offre un bon équilibre entre simplicité et performance. En revanche, FedSGD montre des performances très faibles, avec seulement 10,8% de précision pour des taux d'apprentissage de 0,01 et 0,1 respectivement.



FedProx affiche des résultats prometteurs avec 90% de précision pour des valeurs de μ de 0,01 et 0,1 respectivement. Ces résultats légèrement supérieurs à FedAvg suggèrent que le terme de régularisation proximal aide à stabiliser l'apprentissage.

Les visualisations de progression des paramètres montrent des différences significatives dans les trajectoires d'apprentissage de ces algorithmes:


Progression de l'entraînement fedprox - FedProx_mu0.01
Progression
Round 1
Round 2
Round 3
Round 4
Round 5
Clients
Client 0
Client 1
Client 2
Modèle central
PC1 (87.50% variance)
PC2 (9.35% variance)
0.002
0.000
−0.002
−0.004
−0.020
−0.015
−0.010
−0.005
0.000
0.005



FedAvg et FedProx convergent vers des régions similaires de l'espace des paramètres, tandis que FedSGD semble stagner.

Conclusion et perspectives

Notre étude a permis d'identifier plusieurs facteurs clés influençant l'efficacité de l'apprentissage fédéré. Si cette approche n'atteint pas les performances d'un modèle centralisé ayant accès à toutes les données, elle offre néanmoins un compromis intéressant entre confidentialité et précision.

Les défis majeurs concernent la gestion de l'hétérogénéité des données entre clients et l'équilibre optimal entre le nombre de rounds globaux et d'époques locales. L'algorithme d'agrégation joue également un rôle crucial, avec FedProx démontrant une légère supériorité sur FedAvg dans nos tests.

Ces observations fournissent des pistes pour l'optimisation des systèmes d'apprentissage fédéré dans des applications réelles, où les contraintes de ressources, de confidentialité et de performance doivent être équilibrées. Notre étude démontre qu'avec une configuration appropriée, l'apprentissage fédéré peut atteindre des performances compétitives tout en préservant la confidentialité des données.

Pour des développements futurs, il serait intéressant d'explorer des méthodes plus robustes pour gérer les données non-IID, des techniques de communication plus efficaces pour réduire la bande passante requise, ainsi que des approches de personnalisation adaptant le modèle global aux spécificités locales sans compromettre la performance globale.

1. Introduction à l'Apprentissage Fédéré et ses Enjeux de Sécurité

L'intelligence artificielle contemporaine a connu une évolution remarquable au cours des dernières années, notamment grâce à l'émergence de nouvelles approches qui redéfinissent la manière dont les modèles d'apprentissage sont développés et déployés. Parmi ces innovations, l'Apprentissage Fédéré (Federated Learning ou FL) se distingue comme un paradigme révolutionnaire qui repense fondamentalement la relation entre les données et les algorithmes. Contrairement aux méthodes traditionnelles d'apprentissage automatique où les données des utilisateurs sont centralisées sur un serveur pour l'entraînement des modèles, l'Apprentissage Fédéré propose une approche décentralisée qui préserve la confidentialité des données.

Tel que présenté dans la littérature scientifique, notamment par McMahan et ses collaborateurs, l'Apprentissage Fédéré représente une nouvelle génération d'intelligence artificielle qui s'appuie sur l'entraînement décentralisé, rapprochant ainsi l'apprentissage des appareils périphériques. Ce paradigme innovant est souvent qualifié de "nouvelle aube de l'IA", bien qu'il demeure encore dans ses phases initiales de développement. Sa pleine adoption reste freinée par diverses préoccupations, particulièrement en matière de sécurité et de confidentialité, qui constituent le cœur de notre analyse.

La philosophie sous-jacente à l'Apprentissage Fédéré pourrait être résumée par cette phrase éloquente : "FL amène le code aux données, plutôt que les données au code". Cette approche répond directement aux problématiques de confidentialité, de propriété et de localité des données. Dans ce cadre, les paramètres du modèle voyagent entre un serveur central et les dispositifs clients, permettant ainsi un entraînement collaboratif sans jamais exposer les données brutes des utilisateurs. Cependant, cette circulation des paramètres introduit de nouvelles surfaces d'attaque et vulnérabilités qui nécessitent une attention particulière.

L'émergence de l'Apprentissage Fédéré s'inscrit dans un contexte sociétal et législatif où la protection des données personnelles devient une préoccupation majeure. Avec l'entrée en vigueur de réglementations strictes comme le Règlement Général sur la Protection des Données (RGPD) en Europe et le California Consumer Privacy Act (CCPA) aux États-Unis, les organisations sont confrontées à des exigences croissantes en matière de minimisation des données et de transparence. L'Apprentissage Fédéré apparaît ainsi comme une réponse technologique à ces contraintes réglementaires, permettant l'exploitation des données à des fins d'apprentissage tout en respectant les principes fondamentaux de la protection de la vie privée.

Si la promesse de l'Apprentissage Fédéré repose sur l'amélioration de la confidentialité des données, Mothukuri et al. ainsi que Li et al. soulignent dans leurs travaux que cette approche n'est pas exempte de vulnérabilités sécuritaires et de risques pour la vie privée. Le partage des paramètres du modèle, bien que moins sensible que les données brutes, peut néanmoins révéler des informations confidentielles lorsqu'il est soumis à des techniques d'attaque sophistiquées. De plus, la nature distribuée du processus d'apprentissage ouvre la porte à des altérations malveillantes par des participants adverses. Comprendre ces risques et développer des contre-mesures efficaces constitue un prérequis essentiel pour que l'Apprentissage Fédéré puisse tenir ses promesses en matière de protection des données sans compromettre la performance des modèles générés.

2. Paysage des Vulnérabilités et Menaces

L'Apprentissage Fédéré s'articule autour de plusieurs approches et techniques dont les implications en matière de sécurité et de confidentialité varient considérablement. Cette section explore systématiquement les vulnérabilités sécuritaires inhérentes aux différentes architectures et protocoles de FL, offrant ainsi une cartographie complète des risques potentiels.

Modèles d'Attaque et Classification des Menaces

Avant d'examiner les vulnérabilités spécifiques, il est essentiel d'établir une taxonomie structurée des modèles d'attaque dans le contexte de l'Apprentissage Fédéré. Mothukuri et al. proposent une classification selon plusieurs dimensions clés.

La première dimension concerne le rôle de l'attaquant dans le système fédéré. Un attaquant peut agir en tant que participant au processus d'apprentissage (client malveillant), en tant que serveur d'agrégation, ou comme observateur externe interceptant les communications. Chaque position offre des capacités et contraintes distinctes pour l'adversaire.

Une deuxième dimension fondamentale est l'objectif de l'attaque. Les attaques peuvent viser à compromettre l'intégrité du modèle (en dégradant ses performances ou en induisant des comportements spécifiques), à extraire des informations confidentielles des données d'entraînement, ou à perturber la disponibilité du système (par exemple, en causant une divergence dans le processus d'apprentissage).

La connaissance préalable à disposition de l'attaquant constitue une troisième dimension critique. Les attaques en boîte blanche supposent une connaissance complète de l'architecture du modèle et parfois des données d'autres participants, tandis que les attaques en boîte noire s'appuient uniquement sur les informations publiques ou accessibles via des interfaces légitimes.

Vulnérabilités dans les Architectures Fédérées

Les différentes architectures de l'Apprentissage Fédéré présentent des profils de vulnérabilité distincts. Dans l'architecture centralisée classique (client-serveur), le serveur d'agrégation représente un point unique de défaillance, dont la compromission peut affecter l'ensemble du système. De plus, ce serveur central dispose d'une visibilité privilégiée sur les mises à jour de tous les participants, ce qui peut faciliter les attaques par inférence si ce serveur est malveillant ou compromis.

Les architectures décentralisées (pair-à-pair) éliminent cette vulnérabilité centrale mais introduisent de nouveaux risques. L'absence d'entité de coordination complique l'identification et l'exclusion des nœuds malveillants. De plus, les communications directes entre participants multiplient les surfaces d'attaque et peuvent faciliter les attaques ciblées ou l'établissement de coalitions adverses.

Les architectures hiérarchiques, qui intègrent des agrégateurs intermédiaires, présentent quant à elles des vulnérabilités hybrides. La compromission d'un agrégateur intermédiaire peut affecter tout un sous-ensemble de participants sans nécessairement compromettre l'ensemble du système. Cette architecture peut également faciliter les attaques sélectives, où un adversaire cible spécifiquement certains groupes de participants.

Protocoles de Communication et Vulnérabilités

L'aspect communication constitue une dimension critique de la sécurité dans l'Apprentissage Fédéré. Les protocoles synchrones, qui attendent la contribution de tous les participants sélectionnés avant de mettre à jour le modèle global, sont particulièrement vulnérables aux attaques de déni de service. Un attaquant peut retarder considérablement le processus d'apprentissage en ralentissant délibérément ses propres contributions.

Les protocoles asynchrones, bien que plus résistants aux retards intentionnels, introduisent d'autres vulnérabilités. La possibilité d'intégrer les contributions au fur et à mesure de leur disponibilité peut faciliter les attaques par empoisonnement ciblées temporellement, où un adversaire attend que certaines conditions soient réunies avant de soumettre une contribution malveillante.

Les mécanismes de sélection des participants présentent également des risques significatifs. Si la sélection est prévisible ou manipulable, un adversaire pourrait augmenter stratégiquement sa présence dans certaines rounds d'entraînement pour maximiser son impact sur le modèle global.

Frameworks et Implémentations: Analyse de Sécurité

L'examen des frameworks d'Apprentissage Fédéré les plus utilisés révèle des profils de sécurité variables. TensorFlow Federated (TFF) propose des mécanismes de sécurité comme l'agrégation sécurisée et des options pour la confidentialité différentielle, mais ces fonctionnalités restent optionnelles et leur activation peut significativement impacter les performances. FATE intègre davantage de mécanismes cryptographiques par défaut, notamment pour les scénarios d'Apprentissage Fédéré vertical, mais impose en contrepartie une surcharge computationnelle plus importante.

PySyft se distingue par son intégration poussée de primitives cryptographiques comme le calcul multi-parties sécurisé et le chiffrement homomorphe, offrant des garanties théoriques fortes mais au prix d'une complexité accrue et de limitations pratiques significatives en termes de performance. Flower, plus récent, met l'accent sur la flexibilité mais laisse largement la responsabilité des mécanismes de sécurité aux développeurs.

Cette diversité d'approches souligne l'absence actuelle de standards de sécurité établis dans le domaine de l'Apprentissage Fédéré, laissant aux implémenteurs la charge de naviguer entre garanties théoriques, contraintes pratiques et exigences spécifiques à leur contexte d'application.

3. Analyse Approfondie des Menaces de Sécurité

L'Apprentissage Fédéré, tout en résolvant certains problèmes traditionnels de confidentialité liés à la centralisation des données, introduit un ensemble de vulnérabilités spécifiques qui méritent une analyse approfondie. Les travaux de Mothukuri et al. ainsi que de Li et al. proposent une cartographie exhaustive de ces menaces, que nous pouvons organiser selon leur nature fondamentale : les attaques visant l'intégrité du modèle, celles ciblant la confidentialité des données, et celles affectant la disponibilité du système.

Attaques contre l'Intégrité du Modèle

Empoisonnement des Données

L'empoisonnement des données (*data poisoning*) représente l'une des menaces les plus directes contre l'intégrité des modèles fédérés. Dans sa forme la plus simple, cette attaque consiste pour un participant malveillant à manipuler délibérément ses données d'entraînement locales afin d'induire des comportements erronés dans le modèle global. Cette manipulation peut prendre plusieurs formes :

L'empoisonnement non ciblé (*untargeted poisoning*) vise à dégrader globalement les performances du modèle, rendant ses prédictions généralement inexactes. Cette approche est relativement simple à mettre en œuvre mais également plus facile à détecter, car elle provoque une dégradation importante des métriques de performance.

L'empoisonnement ciblé (*targeted poisoning*) est plus sophistiqué et vise à modifier le comportement du modèle uniquement pour certaines entrées spécifiques, tout en maintenant des performances normales sur le reste des données. Par exemple, un attaquant pourrait tenter de faire mal classifier uniquement une catégorie particulière d'images ou de faire échouer le modèle uniquement sur des cas aux caractéristiques précises.

Les attaques de biais (*bias attacks*) constituent une variante particulièrement préoccupante de l'empoisonnement, visant à introduire ou amplifier des biais discriminatoires dans le modèle. Ces attaques peuvent avoir des implications éthiques et légales significatives, particulièrement dans des domaines sensibles comme l'octroi de crédit, le recrutement ou les décisions médicales.

Empoisonnement des Modèles

L'empoisonnement des modèles (*model poisoning*) représente une forme d'attaque plus directe et potentiellement plus dévastatrice. Plutôt que de manipuler ses données d'entraînement, l'attaquant modifie directement les mises à jour des paramètres de son modèle local avant de les soumettre au serveur d'agrégation. Mothukuri et al. identifient plusieurs variantes de cette attaque :

L'empoisonnement brutal (*scaling attack*) consiste à multiplier les gradients légitimes par un facteur important pour amplifier artificiellement l'influence de l'attaquant sur le modèle global. Cette approche peut être particulièrement efficace dans les systèmes utilisant des agrégateurs simples comme la moyenne non pondérée (FedAvg).

L'empoisonnement sournois (*stealthy poisoning*) est plus sophistiqué et vise à rester indétectable en maintenant les mises à jour du modèle dans une plage statistiquement plausible. Ces attaques exploitent souvent la connaissance des mécanismes de défense déployés pour les contourner délibérément, par exemple en restant sous les seuils de détection des méthodes basées sur des statistiques.

La technique du cheval de Troie neuronale (*neural trojan*) implante des comportements dormants dans le modèle, qui ne se manifestent que lorsque des entrées contenant certains déclencheurs (*triggers*) spécifiques sont présentées. Ces attaques sont particulièrement difficiles à détecter car le modèle se comporte normalement dans la vaste majorité des cas.

Attaques par Porte Dérobée

Les attaques par porte dérobée (*backdoor attacks*) représentent une menace particulièrement sévère pour l'intégrité des modèles fédérés. Contrairement à l'empoisonnement général qui vise à dégrader les performances, ces attaques implantent des vulnérabilités spécifiques que l'attaquant pourra exploiter ultérieurement.

Li et al. décrivent plusieurs variantes de ces attaques. Les portes dérobées sémantiques (*semantic backdoors*) exploitent des caractéristiques naturellement présentes dans les données, comme la présence de lunettes sur un visage, pour déclencher le comportement malveillant. Ces attaques sont particulièrement difficiles à détecter car elles n'introduisent pas d'éléments visiblement anormaux dans les données.

Les portes dérobées avec déclencheurs (*trigger-based backdoors*) utilisent des motifs spécifiques ajoutés aux entrées pour activer le comportement malveillant. Dans le contexte de l'Apprentissage Fédéré, ces attaques sont amplifiées par la possibilité pour un attaquant de cibler ses modifications sur plusieurs rounds d'entraînement et d'adapter sa stratégie en fonction de l'évolution du modèle global.

La persistance de ces vulnérabilités constitue un défi majeur. Mothukuri et al. notent que même après l'exclusion d'un client malveillant, les portes dérobées implantées peuvent persister dans le modèle global pendant plusieurs itérations d'entraînement, particulièrement si elles ont été conçues pour résister aux processus d'agrégation et de mise à jour ultérieurs.

Attaques contre la Confidentialité des Données

Attaques par Inférence de Membres

Les attaques par inférence de membres (*membership inference attacks*) visent à déterminer si un échantillon spécifique a fait partie des données d'entraînement d'un participant. Dans le contexte de l'Apprentissage Fédéré, ces attaques peuvent être particulièrement puissantes lorsqu'elles sont menées par le serveur d'agrégation ou par des participants collusifs qui observent les mises à jour des modèles sur plusieurs rounds.

Li et al. décrivent plusieurs techniques d'inférence de membres adaptées au contexte fédéré. L'approche par divergence de confiance (*confidence skew*) exploite les différences de comportement du modèle face à des exemples vus pendant l'entraînement versus des exemples inconnus. Les modèles d'apprentissage automatique ont tendance à produire des prédictions avec une confiance plus élevée pour les échantillons sur lesquels ils ont été entraînés, ce qui crée une signature statistique détectable.

Les attaques par gradient (*gradient-based attacks*) exploitent directement les mises à jour partagées dans le protocole fédéré. En analysant minutieusement les changements dans les gradients à travers plusieurs itérations, un attaquant peut inférer des informations sur les données sous-jacentes, particulièrement pour des échantillons rares ou distinctifs qui influencent fortement certaines parties du modèle.

Ces attaques posent un risque particulier dans des contextes où la simple appartenance à un ensemble de données constitue une information sensible. Par exemple, dans un système fédéré d'analyse médicale, déterminer qu'un individu a contribué à l'entraînement d'un modèle spécifique à une pathologie peut révéler des informations médicales confidentielles.

Attaques par Reconstruction de Données

Les attaques par reconstruction de données (*data reconstruction attacks*) représentent une menace encore plus sévère, visant à reconstruire partiellement ou complètement les données d'entraînement originales à partir des mises à jour du modèle partagées. Ces attaques exploitent le fait que les gradients partagés dans l'Apprentissage Fédéré contiennent intrinsèquement des informations sur les données ayant servi à les calculer.

Mothukuri et al. décrivent plusieurs techniques sophistiquées de reconstruction, notamment les attaques par inversion de gradient (*gradient inversion*) qui tentent de résoudre un problème d'optimisation inverse pour retrouver les entrées qui auraient généré les gradients observés. Dans certains cas, particulièrement avec des modèles surparamétrés comme les réseaux de neurones profonds, ces techniques peuvent reconstruire avec une précision surprenante les exemples d'entraînement.

Les attaques par génération (*generative attacks*) utilisent quant à elles des modèles génératifs auxiliaires, comme les GAN (Generative Adversarial Networks), pour produire des exemples synthétiques qui auraient pu générer les gradients observés. Ces attaques sont particulièrement efficaces lorsque l'attaquant dispose d'informations préalables sur la distribution des données.

La vulnérabilité à ces attaques varie considérablement selon les types de données. Les données structurées hautement dimensionnelles comme les images sont généralement plus vulnérables que les données tabulaires de faible dimension. De même, les modèles plus complexes avec de nombreux paramètres tendent à "mémoriser" davantage leurs données d'entraînement, facilitant ainsi leur reconstruction.

Attaques par Inférence de Propriétés

Les attaques par inférence de propriétés (*property inference attacks*) visent à extraire des informations statistiques sur les données d'entraînement qui n'ont pas d'impact direct sur la tâche d'apprentissage principale. Par exemple, dans un modèle de classification d'images faciales, un attaquant pourrait tenter d'inférer des statistiques démographiques sur la population représentée dans les données d'entraînement d'un participant.

Li et al. soulignent que ces attaques peuvent révéler des informations sensibles même lorsque les données individuelles restent protégées. Dans un contexte d'Apprentissage Fédéré inter-organisationnel, elles pourraient permettre d'extraire des informations commercialement sensibles, comme la composition du portefeuille client d'une banque participant à un consortium sur la détection de fraudes.

Ces attaques exploitent souvent des modèles "shadow" ou auxiliaires, entraînés par l'attaquant pour imiter le comportement du modèle cible lorsqu'il est entraîné sur des données présentant certaines

propriétés. En comparant les motifs de mise à jour entre ces modèles shadows et le modèle ciblé, l'attaquant peut inférer la présence de ces propriétés dans les données d'entraînement de la victime.

Attaques contre la Disponibilité

Attaques par Déni de Service

Les attaques par déni de service (*Denial of Service, DoS*) dans le contexte de l'Apprentissage Fédéré visent à perturber le processus d'entraînement collaboratif en le ralentissant considérablement ou en le rendant impossible. Ces attaques peuvent cibler l'infrastructure de communication ou exploiter les spécificités du protocole fédéré lui-même.

Mothukuri et al. identifient plusieurs variantes de ces attaques. Les attaques par retardement (*straggler attacks*) exploitent les protocoles synchrones en délibérément retardant les contributions de certains participants, forçant l'ensemble du système à attendre. Dans les systèmes impliquant de nombreux participants, même une petite fraction de clients malveillants peut causer des retards significatifs.

Les attaques par surcharge (*flooding attacks*) visent à submerger le serveur d'agrégation ou les canaux de communication avec un volume excessif de mises à jour, potentiellement mal formées ou invalides. Ces attaques peuvent être particulièrement efficaces contre les systèmes disposant de ressources limitées ou ne mettant pas en œuvre des mécanismes robustes de validation et de limitation de débit.

Les attaques par déconnexion sélective (*selective disconnection*) exploitent les mécanismes de tolérance aux pannes des systèmes fédérés. En se déconnectant stratégiquement à certains moments critiques du processus d'apprentissage, des clients malveillants peuvent forcer le système à reprendre certaines étapes ou à reconfigurer le processus d'agrégation, causant des inefficacités significatives.

Attaques de Sybil

Les attaques de Sybil, nommées d'après un cas célèbre de trouble dissociatif de l'identité, impliquent qu'un adversaire crée et contrôle de multiples identités fictives pour augmenter son influence sur le système. Dans le contexte de l'Apprentissage Fédéré, ces attaques peuvent significativement amplifier l'impact des autres types d'attaques précédemment décrits.

Li et al. soulignent que ces attaques sont particulièrement problématiques dans les systèmes fédérés ouverts ou faiblement authentifiés. En contrôlant une proportion significative des participants apparents, un attaquant peut influencer de manière disproportionnée le processus d'agrégation, faciliter des attaques par collusion, ou manipuler les mécanismes de détection d'anomalies basés sur des consensus majoritaires.

Ces attaques posent un défi fondamental car elles remettent en question l'hypothèse implicite que chaque participant représente une entité distincte avec des intérêts divergents. Les mécanismes traditionnels de défense contre les attaques de Sybil, comme les systèmes de réputation ou les

preuves d'identité, peuvent être difficiles à implémenter dans des contextes où l'anonymat ou la pseudonymité des participants est souhaitable pour des raisons de confidentialité.

4. Mécanismes de Défense et Techniques de Protection

Face aux multiples vulnérabilités identifiées, différentes stratégies de défense ont été proposées pour renforcer la sécurité et la confidentialité des systèmes d'Apprentissage Fédéré. Cette section examine les mécanismes actuels et émergents, en analysant leur efficacité, leurs limitations, et les compromis qu'ils imposent entre protection, performance et efficacité computationnelle.

Défenses contre les Attaques d'Intégrité

Mécanismes d'Agrégation Robuste

L'agrégation robuste constitue une première ligne de défense contre les attaques par empoisonnement. Ces approches visent à limiter l'influence des mises à jour malveillantes sur le modèle global en adaptant l'algorithme d'agrégation lui-même. Mothukuri et al. présentent plusieurs variantes de ces mécanismes.

La médiane coordonnée-par-coordonnée (*coordinate-wise median*) remplace la moyenne traditionnelle des gradients par leur médiane, réduisant significativement l'impact des valeurs aberrantes introduites par les attaquants. Cette approche simple offre une robustesse considérable contre les attaques d'empoisonnement brutal, mais peut réduire l'efficacité de l'apprentissage en présence de distributions de données légitimement hétérogènes.

L'agrégation par moyennes tronquées (*trimmed mean*) exclut une fraction des valeurs extrêmes avant de calculer la moyenne des contributions restantes. Le défi réside dans la détermination du seuil de troncature optimal : trop strict, il risque d'éliminer des contributions légitimes mais atypiques ; trop permissif, il laisse passer des contributions malveillantes.

L'agrégation par moyenne pondérée par la réputation (*reputation-weighted aggregation*) attribue des poids différents aux participants en fonction de leur historique de contributions. Ces poids évoluent dynamiquement au cours du processus d'apprentissage, réduisant progressivement l'influence des clients produisant des mises à jour statistiquement aberrantes ou incohérentes temporellement.

Détection d'Anomalies et Filtrage

Les approches basées sur la détection d'anomalies visent à identifier et exclure les contributions malveillantes avant leur intégration dans le modèle global. Li et al. distinguent plusieurs catégories de ces mécanismes.

Les méthodes basées sur les normes (*norm-based detection*) utilisent des statistiques simples comme la norme L1 ou L2 des mises à jour pour identifier les contributions anormalement grandes ou divergentes. Bien que computationnellement efficaces, ces approches peuvent être facilement contournées par des attaquants sophistiqués qui calibrent soigneusement l'amplitude de leurs manipulations.

Les approches basées sur la cohérence (*consistency-based detection*) vérifient la cohérence interne des mises à jour soumises, par exemple en analysant les relations entre différentes couches d'un réseau de neurones. Ces méthodes peuvent détecter des attaques sophistiquées qui maintiennent des statistiques globales plausibles tout en introduisant des incohérences localisées.

La validation croisée fédérée (*federated cross-validation*) consiste à évaluer l'impact des mises à jour proposées sur un ensemble de validation détenu par le serveur ou partagé entre les participants honnêtes. Cette approche peut identifier efficacement les contributions qui dégradent les performances globales, mais nécessite des données de validation représentatives et soulève des questions de confidentialité supplémentaires.

Certification et Défenses Provables

Les défenses certifiées (*certified defenses*) visent à fournir des garanties mathématiques formelles quant à la robustesse du modèle face à certaines classes d'attaques. Ces approches, encore émergentes dans le contexte fédéré, s'inspirent des avancées récentes dans le domaine de l'apprentissage robuste centralisé.

La randomisation certifiée (*randomized smoothing*) transforme un classificateur standard en un classificateur robuste en introduisant un bruit aléatoire contrôlé pendant l'inférence. Dans le contexte fédéré, cette technique peut être adaptée pour offrir des garanties sur la robustesse du modèle global face à des perturbations localisées introduites par des clients malveillants.

Les approches par enveloppe convexe (*convex envelope methods*) définissent des régions "sûres" dans l'espace des paramètres et contraignent l'agrégation à produire un modèle global restant dans ces régions. Ces méthodes peuvent offrir des garanties formelles contre certaines formes d'empoisonnement, mais au prix d'une complexité accrue et de potentielles limitations en termes de capacité d'apprentissage.

Protections de la Confidentialité

Confidentialité Différentielle

La confidentialité différentielle (DP) constitue l'approche la plus établie pour protéger formellement les données privées dans les systèmes d'Apprentissage Fédéré. Ce cadre mathématique offre des garanties quantifiables quant à la capacité d'un adversaire à inférer des informations sur les données individuelles à partir des sorties du système.

Li et al. détaillent plusieurs variantes de la DP adaptées au contexte fédéré. La confidentialité différentielle centralisée (*central DP*) considère le serveur d'agrégation comme une entité de confiance et vise à protéger les données contre les attaques externes. Cette approche permet généralement d'atteindre un meilleur compromis utilité-confidentialité, mais repose sur l'hypothèse forte d'un serveur non malveillant.

La confidentialité différentielle locale (*local DP*) offre des garanties plus fortes en supposant que même le serveur d'agrégation peut être malveillant. Chaque client perturbe ses mises à jour localement avant de les partager, assurant ainsi que même le serveur ne peut extraire d'informations

sensibles. Cette approche impose cependant un coût significatif en termes de qualité du modèle final, particulièrement dans les scénarios impliquant de nombreux participants.

Des extensions comme la confidentialité différentielle répartie (*distributed DP*) tentent d'optimiser ce compromis en partageant la responsabilité de l'ajout de bruit entre les clients et le serveur. Ces approches hybrides nécessitent généralement des primitives cryptographiques supplémentaires pour garantir que le bruit total satisfait les exigences de confidentialité différentielle sans qu'aucune entité individuelle ne puisse le contourner.

Techniques Cryptographiques

Les approches cryptographiques visent à protéger la confidentialité des données et des modèles par des garanties mathématiques fortes, sans nécessairement introduire de dégradation intentionnelle de l'information comme le fait la confidentialité différentielle.

Le calcul multi-parties sécurisé (SMC) permet à plusieurs entités de calculer conjointement une fonction sur leurs données privées sans révéler ces données aux autres participants. Dans le contexte de l'Apprentissage Fédéré, Mothukuri et al. décrivent comment le SMC peut être utilisé pour réaliser l'agrégation des mises à jour de modèles de manière confidentielle, empêchant même le serveur d'agrégation d'accéder aux contributions individuelles. Ces protocoles reposent généralement sur des techniques comme le partage de secrets ou les circuits garblés, qui imposent une surcharge significative en termes de communication et de calcul.

Le chiffrement homomorphe permet d'effectuer des opérations mathématiques directement sur des données chiffrées, sans nécessiter leur déchiffrement préalable. Cette propriété remarquable permet théoriquement de réaliser l'ensemble du processus d'Apprentissage Fédéré sur des modèles et des mises à jour chiffrés. Toutefois, les schémas de chiffrement homomorphe actuels imposent une surcharge computationnelle prohibitive, rendant leur application pratique limitée à des opérations spécifiques comme l'agrégation des mises à jour plutôt qu'à l'ensemble du processus d'apprentissage.

Les preuves à divulgation nulle de connaissance (*zero-knowledge proofs*) permettent à un participant de prouver qu'il a correctement exécuté certaines opérations sans révéler aucune information supplémentaire. Dans le contexte fédéré, ces techniques peuvent être utilisées pour garantir l'intégrité des mises à jour soumises (par exemple, prouver qu'elles résultent d'un processus d'entraînement légitime sur des données valides) sans compromettre la confidentialité.

Techniques d'Extraction Minimale d'Information

Une approche complémentaire consiste à minimiser intrinsèquement la fuite d'information dans les mises à jour partagées, indépendamment des protections cryptographiques ou différentielles appliquées.

La distillation des connaissances (*knowledge distillation*) permet de partager uniquement les prédictions d'un modèle local sur un ensemble de données publiques plutôt que les paramètres du modèle lui-même. Cette approche limite considérablement les informations transmises, réduisant ainsi la surface d'attaque pour les inférences de données privées. Li et al. notent toutefois que cette

technique peut significativement ralentir la convergence et nécessite un ensemble de données publiques représentatif.

L'élagage et la quantification (*pruning and quantization*) des mises à jour réduisent la granularité et la dimensionnalité de l'information partagée. En ne transmettant qu'une version compressée et simplifiée des mises à jour du modèle, ces techniques limitent naturellement la capacité d'un adversaire à extraire des informations précises sur les données d'entraînement. Ces approches présentent l'avantage supplémentaire de réduire les coûts de communication, un facteur critique dans les déploiements à grande échelle.

Compromis Fondamentaux et Limites Théoriques

Les mécanismes de défense présentés illustrent le trilemme fondamental de l'Apprentissage Fédéré identifié par Li et al. : il semble impossible d'optimiser simultanément la confidentialité, l'efficacité computationnelle et la précision du modèle. Chaque technique de protection impose des compromis spécifiques entre ces dimensions.

La confidentialité différentielle dégrade intentionnellement la qualité des mises à jour pour garantir la protection des données individuelles. Plus le niveau de protection souhaité est élevé (exprimé par un paramètre ε plus petit), plus la dégradation des performances du modèle est importante.

Les approches cryptographiques préservent théoriquement la précision du modèle mais imposent une surcharge computationnelle et communicationnelle considérable, compromettant l'efficacité du système, particulièrement dans des contextes aux ressources limitées comme les appareils mobiles ou l'IoT.

Les mécanismes d'agrégation robuste, en filtrant ou atténuant certaines contributions pour protéger l'intégrité du modèle, risquent simultanément d'écarter des informations légitimes mais atypiques, affectant potentiellement la capacité du modèle à capturer des patterns rares mais importants dans les données.

Ces compromis incontournables soulignent l'importance d'une approche contextuelle de la sécurité et de la confidentialité dans l'Apprentissage Fédéré. Plutôt qu'une solution universelle, les concepteurs de systèmes doivent sélectionner et calibrer soigneusement les mécanismes de protection en fonction des menaces spécifiques, des contraintes de ressources et des exigences de performance propres à chaque application.

5. Perspectives de Recherche et Innovations

Face aux vulnérabilités identifiées et aux limitations des mécanismes de défense actuels, plusieurs pistes de recherche émergent pour renforcer la sécurité et la confidentialité des systèmes d'Apprentissage Fédéré. Ces orientations futures, s'inspirant des travaux de Mothukuri et al. et Li et al., cherchent à résoudre le trilemme fondamental entre confidentialité, performance et efficacité computationnelle.

Architectures de Sécurité Avancées

Architectures Multi-niveaux Sécurisées

Une direction prometteuse concerne le développement d'architectures de sécurité multi-niveaux qui intègrent différentes approches de protection à différentes étapes du processus fédéré. Ces architectures pourraient combiner des mécanismes complémentaires pour atténuer un large spectre de menaces tout en optimisant le compromis entre sécurité et performance.

Les chercheurs proposent des systèmes où la confidentialité différentielle serait appliquée aux données particulièrement sensibles, tandis que des techniques cryptographiques plus lourdes seraient réservées aux opérations critiques comme l'agrégation finale. Cette stratification permettrait d'adapter dynamiquement le niveau de protection en fonction de la sensibilité des données et des opérations, plutôt que d'imposer un niveau uniforme à l'ensemble du système.

Sécurité Adaptative Contextuelle

Les systèmes de sécurité adaptatifs, qui ajustent leurs mécanismes de protection en fonction du contexte d'exécution, offrent une voie prometteuse pour optimiser l'équilibre entre sécurité et performance. Ces systèmes pourraient moduler leurs paramètres de protection en fonction de multiples facteurs:

L'analyse comportementale des participants permettrait d'ajuster le niveau de confiance accordé à chacun en fonction de son historique de contributions. Des participants ayant démontré une fiabilité constante pourraient bénéficier de contrôles allégés, libérant ainsi des ressources pour une surveillance plus stricte des nouveaux participants ou de ceux présentant des comportements inhabituels.

L'adaptation aux menaces détectées permettrait au système de renforcer spécifiquement les défenses contre les types d'attaques identifiés comme actifs ou probables. Par exemple, face à des signes d'attaques par empoisonnement, le système pourrait temporairement privilégier des mécanismes d'agrégation plus robustes, même au prix d'une convergence plus lente.

L'hétérogénéité des mécanismes de défense constitue en elle-même une protection, en compliquant la tâche d'un attaquant cherchant à contourner des défenses prévisibles. Des systèmes incorporant une certaine randomisation dans leurs stratégies de défense pourraient significativement renforcer leur résilience face à des attaquants adaptifs.

Intégration Cryptographique Avancée

Cryptographie Post-Quantique

Anticipant les avancées en informatique quantique qui pourraient compromettre les primitives cryptographiques actuelles, l'intégration de techniques cryptographiques post-quantiques dans les protocoles d'Apprentissage Fédéré représente une direction de recherche cruciale. Ces techniques garantiraient la sécurité à long terme des données et des modèles, même face à des adversaires disposant de capacités de calcul quantique.

Les réseaux de hachage (*hash-based signatures*) et les cryptosystèmes basés sur les réseaux (*lattice-based cryptography*) figurent parmi les candidats les plus prometteurs pour remplacer les primitives actuelles dans le contexte fédéré. L'adaptation efficace de ces techniques aux contraintes spécifiques

de l'Apprentissage Fédéré, particulièrement en termes d'efficacité computationnelle, constitue un défi majeur pour les années à venir.

Chiffrement Homomorphe Optimisé

Les récentes avancées en chiffrement homomorphe, permettant des calculs directs sur des données chiffrées, ouvrent des perspectives intéressantes pour l'Apprentissage Fédéré. Bien que les schémas généraux restent trop coûteux pour une application pratique, des schémas partiellement homomorphes optimisés pour certaines opérations spécifiques pourraient offrir un compromis viable.

Des recherches prometteuses se concentrent sur le développement de primitives homomorphes "sur mesure" pour les opérations d'agrégation fédérée, significativement plus efficaces que les approches génériques. Ces schémas exploitent les particularités mathématiques des opérations d'agrégation (principalement des additions pondérées) pour réduire drastiquement la surcharge computationnelle tout en maintenant de fortes garanties de confidentialité.

L'approche du chiffrement homomorphe seuillé (*threshold homomorphic encryption*), où la clé de déchiffrement est partagée entre plusieurs entités nécessitant leur collaboration pour déchiffrer, offre également des perspectives intéressantes pour les scénarios inter-organisationnels, où aucune entité individuelle ne devrait pouvoir accéder aux contributions brutes.

Techniques de Détection et Réponse Avancées

Détection Basée sur l'Intelligence Artificielle

L'application de techniques avancées d'intelligence artificielle pour la détection des comportements malveillants représente une piste de recherche particulièrement prometteuse. Ces approches exploitent des modèles sophistiqués pour identifier des patterns subtils dans les contributions des participants, potentiellement indétectables par des méthodes statistiques traditionnelles.

Les réseaux adverses génératifs (GANs) peuvent être adaptés pour créer des "détecteurs d'anomalies" spécialisés dans l'identification des mises à jour de modèle suspectes. Ces systèmes seraient entraînés sur des exemples de contributions légitimes et malveillantes pour développer une compréhension nuancée des caractéristiques distinctives des attaques.

Les techniques d'apprentissage par renforcement permettraient de développer des agents de détection qui s'améliorent continuellement au fur et à mesure de leur exposition à de nouvelles variantes d'attaques. Cette adaptabilité dynamique est cruciale dans un contexte où les stratégies d'attaque évoluent constamment pour contourner les défenses existantes.

Les approches d'apprentissage automatique explicable (*explainable ML*) sont particulièrement pertinentes dans ce contexte, car elles permettraient non seulement d'identifier les contributions potentiellement malveillantes, mais également de fournir des justifications compréhensibles pour ces décisions, facilitant ainsi la validation humaine et l'amélioration continue des systèmes de détection.

Réponse Automatisée aux Incidents

Au-delà de la simple détection, le développement de mécanismes automatisés de réponse aux incidents de sécurité constitue une direction de recherche essentielle. Ces systèmes viseraient à maintenir l'intégrité et les performances du processus d'apprentissage même en présence d'attaques actives.

Les techniques de récupération sélective (*selective recovery*) permettraient de "réparer" un modèle compromis sans nécessiter un réentraînement complet. En isolant les composants du modèle potentiellement affectés par des contributions malveillantes, ces approches pourraient significativement réduire l'impact des attaques réussies.

Les mécanismes d'isolation dynamique des participants (*dynamic participant isolation*) permettraient d'exclure temporairement les clients présentant des comportements suspects, tout en leur offrant des voies de réhabilitation après vérification ou correction. Cette approche maintient l'ouverture du système tout en limitant les risques associés à des participants compromis.

Les stratégies de diversification et redondance (*diversity and redundancy*) s'inspirent des principes de tolérance aux pannes en systèmes distribués. En maintenant plusieurs versions du modèle entraînées sur différents sous-ensembles de participants, ces approches pourraient détecter les divergences suspectes et basculer vers les versions non compromises en cas d'attaque détectée.

Cadres Théoriques Unifiés

Théorie Formelle de la Sécurité Fédérée

L'établissement d'un cadre théorique unifié pour la sécurité et la confidentialité dans l'Apprentissage Fédéré représente un défi intellectuel majeur mais essentiel. Ce cadre permettrait de formaliser rigoureusement les propriétés de sécurité souhaitées, les modèles de menace pertinents, et les garanties offertes par différentes approches défensives.

Les chercheurs s'inspirent des fondements théoriques établis dans différents domaines connexes – confidentialité différentielle, cryptographie, théorie des jeux, apprentissage robuste – pour développer un formalisme adapté aux spécificités de l'Apprentissage Fédéré. Ce cadre faciliterait l'analyse comparative des mécanismes de protection existants et guiderait le développement de nouvelles approches.

Un aspect particulièrement crucial de cette théorisation concerne la quantification précise des compromis entre différentes propriétés désirables – confidentialité, robustesse, efficacité, précision. Une meilleure compréhension théorique de ces relations permettrait aux concepteurs de systèmes de faire des choix éclairés adaptés à leur contexte spécifique.

Métriques Standardisées d'Évaluation

Le développement de métriques standardisées pour évaluer la sécurité et la confidentialité des systèmes d'Apprentissage Fédéré constitue un prérequis pour des comparaisons objectives et des avancées systématiques dans ce domaine. Ces métriques devraient couvrir les différentes dimensions de la sécurité fédérée et être adaptées aux spécificités de ce paradigme.

Les mesures de robustesse aux attaques doivent quantifier la résistance du système face à différents types d'attaques, sous diverses contraintes sur les capacités de l'adversaire. Ces métriques devraient aller au-delà des scénarios simplifiés pour capturer des situations réalistes où les attaquants adaptent dynamiquement leurs stratégies.

Les métriques de fuite d'information doivent évaluer quantitativement le risque que des informations sensibles soient déduites des mises à jour partagées ou du modèle final. Ces mesures devraient refléter différents niveaux de connaissance préalable de l'attaquant et divers objectifs d'inférence, des plus généraux aux plus spécifiques.

Les indicateurs d'overhead opérationnel doivent capturer les coûts additionnels imposés par les mécanismes de sécurité en termes de calcul, communication, et dégradation potentielle des performances du modèle. Ces métriques facilitent l'analyse du rapport coût-bénéfice des différentes approches défensives.

Apprentissage Fédéré à Confidentialité Préservée par Construction

Architectures Préservant Intrinsèquement la Confidentialité

Une direction de recherche particulièrement prometteuse concerne le développement d'architectures d'Apprentissage Fédéré qui préservent intrinsèquement la confidentialité, sans nécessiter l'ajout de mécanismes de protection supplémentaires. Ces architectures intégreraient les considérations de confidentialité dès leur conception fondamentale plutôt que comme une couche additionnelle.

Les approches par distillation (*distillation-based approaches*) exploitent le principe que le transfert de connaissances entre modèles peut être réalisé en partageant uniquement des prédictions sur des données publiques, plutôt que les paramètres des modèles eux-mêmes. Cette indirection réduit naturellement la fuite d'information tout en permettant un apprentissage collaboratif efficace.

Les architectures modulaires (*modular architectures*) décomposent le modèle en composants avec différentes sensibilités et exigences de confidentialité. Cette décomposition permet d'appliquer différents niveaux de protection à différents éléments du modèle, optimisant ainsi le compromis global entre confidentialité et performance.

Les approches basées sur les représentations (*representation-based approaches*) visent à identifier et partager uniquement des représentations abstraites des données qui conservent l'information utile à la tâche d'apprentissage tout en éliminant les détails sensibles. Ces techniques s'inspirent des principes de l'apprentissage par transfert et des auto-encodeurs, adaptés au contexte fédéré.

Confidentialité Différentielle Locale Optimisée

Reconnaissant les limitations pratiques de la confidentialité différentielle traditionnelle, particulièrement en termes de dégradation des performances, les chercheurs explorent des variantes optimisées adaptées spécifiquement au contexte fédéré.

Les mécanismes d'allocation adaptative de budget de confidentialité (*adaptive privacy budget allocation*) ajustent dynamiquement le niveau de bruit appliqué en fonction de différents facteurs

contextuels: sensibilité des données traitées, état de convergence du modèle, ou importance relative de différentes dimensions des mises à jour.

Les approches par échantillonnage amplificateur de confidentialité (*privacy amplification by sampling*) exploitent le fait que la sélection aléatoire d'un sous-ensemble de participants ou de paramètres à chaque itération renforce naturellement les garanties de confidentialité. Cette propriété permet de réduire la quantité de bruit explicitement ajoutée tout en maintenant un niveau équivalent de protection.

Les techniques de confidentialité différentielle spécifiques à la tâche (*task-specific differential privacy*) incorporent des connaissances sur la structure et les exigences particulières du problème d'apprentissage pour optimiser l'ajout de bruit. En exploitant les spécificités du domaine, ces approches peuvent préserver la performance sur les aspects critiques de la tâche tout en protégeant efficacement les informations sensibles.

6. Études de Cas: Sécurité dans les Déploiements

L'examen des déploiements concrets d'Apprentissage Fédéré révèle comment les considérations théoriques de sécurité et de confidentialité se traduisent en solutions pratiques. Ces études de cas illustrent les défis spécifiques rencontrés dans différents contextes et les approches adoptées pour les surmonter.

Sécurité dans les Systèmes de Santé Fédérés

Le secteur de la santé, avec ses exigences strictes en matière de confidentialité des données et ses enjeux critiques, offre des exemples particulièrement instructifs de déploiements sécurisés d'Apprentissage Fédéré.

Le Consortium MELLODDY: Protection de la Propriété Intellectuelle

Le projet MELLODDY (Machine Learning Ledger Orchestration for Drug Discovery) représente un cas d'étude remarquable d'Apprentissage Fédéré dans un contexte hautement compétitif. Ce consortium, réunissant dix grandes entreprises pharmaceutiques européennes, vise à développer collaborativement des modèles prédictifs pour la découverte de médicaments sans compromettre leurs secrets industriels respectifs.

L'architecture de sécurité de MELLODDY repose sur plusieurs couches de protection. Au niveau cryptographique, le système utilise une combinaison de chiffrement homomorphe pour les opérations d'agrégation et de calcul multi-parties sécurisé pour certaines phases critiques du processus. Cette approche garantit qu'aucune entreprise participante ne peut accéder aux données brutes ou aux contributions spécifiques des autres participants.

Pour prévenir les attaques par inférence, MELLODDY intègre des mécanismes de confidentialité différentielle spécifiquement calibrés pour les données chimiques. Le niveau de bruit est ajusté dynamiquement en fonction de la sensibilité des différentes caractéristiques moléculaires, offrant une protection renforcée pour les informations les plus stratégiques tout en préservant l'utilité du modèle pour les prédictions communes.

Le système implémente également des contrôles rigoureux contre les attaques par empoisonnement. Chaque contribution est validée par un comité technique indépendant qui vérifie sa conformité avec des critères prédéfinis sans accéder aux données sous-jacentes. De plus, un mécanisme de réputation basé sur la blockchain enregistre de manière immuable l'historique des contributions de chaque participant, créant ainsi une incitation structurelle à la coopération honnête.

Les leçons clés tirées de ce déploiement soulignent l'importance d'une conception de sécurité adaptée au contexte spécifique. Les mécanismes de protection doivent être calibrés en fonction des motivations et capacités probables des adversaires potentiels – dans ce cas, des concurrents industriels sophistiqués avec des intérêts économiques substantiels.

EXAM: Confidentialité des Données Médicales Inter-institutionnelles

L'initiative EXAM (Electronic Cross-Hospital Artificial Intelligence Medicine) illustre les défis de sécurité spécifiques aux collaborations inter-hospitalières. Ce projet vise à développer des modèles de diagnostic basés sur l'imagerie médicale provenant de multiples institutions, tout en respectant les réglementations strictes comme HIPAA aux États-Unis ou le RGPD en Europe.

L'architecture de sécurité d'EXAM se distingue par son approche modulaire. Plutôt que d'appliquer uniformément des mécanismes de protection coûteux à l'ensemble du processus, le système catégorise les différentes opérations selon leur niveau de risque et applique des protections proportionnées. Par exemple, les métadonnées non sensibles (comme les types d'équipement ou les protocoles d'imagerie) sont partagées avec des protections minimales, tandis que les gradients susceptibles de révéler des informations patient bénéficient de protections cryptographiques avancées.

Pour contrer les attaques par reconstruction d'images, particulièrement préoccupantes dans ce contexte, EXAM implémente une combinaison de techniques de minimisation d'information. Les images originales sont prétraitées localement pour extraire uniquement les caractéristiques pertinentes pour le diagnostic, réduisant ainsi drastiquement le risque de reconstruction tout en maintenant la performance diagnostique.

Le système intègre également des mécanismes de vérification d'intégrité spécifiquement adaptés au contexte médical. Des ensembles de validation standardisés, composés de cas cliniques annotés par des experts, permettent d'évaluer l'impact de chaque mise à jour sur les performances diagnostiques critiques. Cette approche permet de détecter rapidement les contributions potentiellement malveillantes qui pourraient compromettre la précision diagnostique.

L'expérience d'EXAM souligne l'importance cruciale d'une approche de sécurité qui préserve la sensibilité clinique des modèles. Les mécanismes de protection ne doivent pas simplement garantir la confidentialité abstraite des données, mais également maintenir la précision diagnostique pour les cas critiques où des erreurs pourraient avoir des conséquences graves pour les patients.

Protection des Données Financières dans les Systèmes Fédérés

Le secteur financier, avec ses exigences réglementaires strictes et les implications potentiellement graves des brèches de sécurité, a développé des approches particulièrement sophistiquées pour

sécuriser les déploiements d'Apprentissage Fédéré.

Détection de Fraude Inter-bancaire: Le Cas du Consortium Européen

Un consortium de banques européennes a déployé une solution d'Apprentissage Fédéré pour la détection de fraudes par carte de crédit, permettant d'identifier des schémas de fraude qui resteraient invisibles dans les données isolées de chaque institution. Ce cas illustre les défis spécifiques de sécurité dans un contexte où les données sont extrêmement sensibles et où les adversaires (fraudeurs) sont hautement motivés et adaptables.

L'architecture de sécurité de ce système repose sur un modèle de confiance sophistiqué. Plutôt qu'une approche binaire (entités de confiance versus entités non fiables), le système implémente un modèle de confiance graduée où différents niveaux d'accès et de privilèges sont attribués en fonction de multiples facteurs: historique de l'institution, volume de données contribué, qualité des contributions précédentes, et contrôles de sécurité interne vérifiés.

Pour contrer les attaques par inférence ciblant des informations commercialement sensibles (comme la composition du portefeuille client d'une banque), le système implémente une combinaison de techniques de confidentialité différentielle et de partitionnement sécurisé. Les mises à jour sont segmentées et agrégées séparément selon différentes dimensions (géographique, démographique, temporelle), empêchant ainsi la déduction de la structure globale du portefeuille tout en permettant l'identification de schémas de fraude.

Face au risque d'empoisonnement stratégique, où un fraudeur pourrait tenter d'affaiblir délibérément la détection de certains types de fraudes, le système emploie une approche d'agrégation robuste multi-niveau. Les contributions sont d'abord agrégées au sein de chaque institution, puis entre institutions de taille comparable, avant une agrégation finale pondérée par des facteurs de confiance. Cette structure limite l'influence maximale que pourrait exercer une entité compromise.

Les leçons tirées de ce déploiement soulignent l'importance d'aligner les mécanismes de sécurité avec les modèles d'incitation économique des participants. Dans ce contexte, toutes les institutions partagent un intérêt fondamental à maintenir l'intégrité du système, créant une base solide pour la coopération malgré leurs positions concurrentielles sur d'autres aspects.

FATE: Sécurisation des Applications de Crédit Fédérées

La plateforme FATE (Federated AI Technology Enabler), déployée dans plusieurs institutions financières chinoises pour l'évaluation des risques de crédit, illustre une approche différente de la sécurité fédérée, adaptée à un contexte où les autorités réglementaires jouent un rôle plus direct.

L'architecture de sécurité de FATE se distingue par son intégration poussée avec les infrastructures réglementaires nationales. Le système implémente un modèle de "vérifiabilité réglementaire" où les autorités de régulation disposent d'interfaces spécifiques leur permettant d'auditer le fonctionnement du système sans compromettre la confidentialité des données sous-jacentes.

Pour protéger contre les attaques visant à extraire des informations sur les critères sensibles d'évaluation de crédit, FATE implémente un mécanisme sophistiqué de chiffrement fonctionnel. Cette approche permet aux participants d'effectuer uniquement certaines opérations spécifiques sur les modèles chiffrés, strictement nécessaires à la tâche d'évaluation de crédit, tout en rendant techniquement impossible l'extraction des règles ou pondérations précises utilisées par chaque institution.

Face aux préoccupations d'équité et de biais, particulièrement critiques dans les décisions de crédit, le système intègre des mécanismes de validation éthique des modèles. Ces mécanismes vérifient automatiquement que les mises à jour proposées n'amplifient pas les disparités de traitement entre différents groupes démographiques, même lorsque ces mises à jour pourraient améliorer la performance prédictive globale.

L'expérience de FATE met en lumière l'importance d'une conception de sécurité qui intègre explicitement les considérations réglementaires et éthiques dès le départ, plutôt que de les traiter comme des contraintes externes à satisfaire a posteriori.

Sécurité des Systèmes Fédérés Mobiles et IoT

Les déploiements d'Apprentissage Fédéré dans le contexte des technologies mobiles et de l'Internet des Objets présentent des défis de sécurité distincts, principalement liés aux contraintes de ressources et à l'échelle massive des déploiements.

Gboard: Protection de la Confidentialité des Données de Frappe

L'implémentation par Google de l'Apprentissage Fédéré pour améliorer son clavier Gboard représente l'un des déploiements les plus massifs de cette technologie, impliquant des millions d'appareils. Ce cas illustre les défis de sécurité spécifiques aux déploiements à très grande échelle sur des appareils personnels aux ressources limitées.

L'architecture de sécurité de Gboard repose sur une approche minimaliste soigneusement optimisée. Plutôt que d'implémenter des mécanismes cryptographiques lourds sur les appareils clients, le système utilise une combinaison de techniques légères: échantillonnage aléatoire des participants, mise à jour partielle des modèles, agrégation différentiellement privée, et suppression des mises à jour statistiquement aberrantes.

Pour contrer spécifiquement les attaques visant à reconstruire le texte tapé par les utilisateurs, le système implémente un mécanisme sophistiqué de filtrage des n-grammes rares. Cette approche identifie et supprime les séquences de caractères suffisamment uniques pour potentiellement identifier un utilisateur ou révéler des informations sensibles, tout en préservant les patterns de frappe communs nécessaires à l'amélioration des suggestions textuelles.

Face au risque d'attaques par modèle inversé, où un adversaire pourrait tenter de reconstruire les données d'entraînement à partir du modèle global, Gboard implémente une stratégie de "dégradation contrôlée". Certains aspects du modèle particulièrement susceptibles de mémoriser des données spécifiques sont délibérément simplifiés, réduisant ainsi le risque de fuite tout en maintenant les fonctionnalités essentielles.

Les leçons tirées de ce déploiement soulignent l'importance cruciale de l'efficacité computationnelle dans les mécanismes de sécurité destinés aux environnements contraints. L'expérience de Gboard démontre qu'une combinaison soigneusement orchestrée de techniques légères peut offrir des garanties de sécurité substantielles sans imposer de surcharge prohibitive aux appareils des utilisateurs.

Systèmes Énergétiques Intelligents: Sécurité des Données Comportementales

Les déploiements d'Apprentissage Fédéré dans les systèmes de gestion énergétique intelligente illustrent les défis spécifiques liés à la protection des données comportementales hautement sensibles dans un contexte d'Internet des Objets.

L'architecture de sécurité de ces systèmes se distingue par son approche de "confidentialité par conception" au niveau des capteurs mêmes. Plutôt que de collecter puis de protéger des données brutes potentiellement invasives, les dispositifs extraient localement uniquement les caractéristiques agrégées pertinentes pour l'optimisation énergétique, rendant techniquement impossible la reconstruction des comportements précis des occupants.

Pour contrer les attaques par inférence temporelle, particulièrement préoccupantes dans ce contexte où les données révèlent des routines quotidiennes, le système implémente des mécanismes sophistiqués d'obscurcissement temporel. Les contributions sont délibérément désynchronisées et agrégées sur des fenêtres temporelles variables, empêchant ainsi la reconstruction précise des moments d'activité dans les habitations.

Face aux risques de manipulation malveillante des prédictions énergétiques, qui pourrait potentiellement déstabiliser les réseaux électriques, le système déploie une approche multicouche de validation des modèles. Les mises à jour sont évaluées non seulement pour leur impact sur la précision prédictive, mais également pour leurs implications potentielles sur la stabilité du réseau et la sécurité des approvisionnements.

L'expérience de ces déploiements souligne l'importance d'une conception de sécurité qui considère les implications systémiques plus larges, au-delà de la simple protection des données individuelles. Dans le contexte des infrastructures critiques comme les réseaux énergétiques, la sécurité doit englober non seulement la confidentialité des données mais également la résilience du système global face à des manipulations potentiellement déstabilisatrices.

Voici la section 7 sur les implications réglementaires et éthiques de l'Apprentissage Fédéré :

7. Implications Réglementaires et Éthiques

Cadres Réglementaires et Conformité

Conformité aux Réglementations de Protection des Données

L'Apprentissage Fédéré émerge dans un paysage réglementaire en pleine évolution concernant la protection des données personnelles. Si cette approche semble naturellement alignée avec l'esprit

des législations comme le RGPD européen ou le CCPA californien, son implémentation pratique soulève néanmoins des questions d'interprétation juridique significatives.

Mothukuri et al. soulignent que malgré l'absence de centralisation directe des données brutes, certains aspects de l'Apprentissage Fédéré pourraient toujours tomber sous le coup de ces réglementations. Les mises à jour de modèles partagées pourraient, dans certaines circonstances, être considérées comme des "données personnelles" au sens juridique si elles permettent potentiellement l'identification d'individus via des attaques par inférence sophistiquées.

La notion de "minimisation des données", principe fondamental du RGPD, trouve une résonance particulière dans le contexte fédéré. Les implémentations qui extraient et partagent uniquement les caractéristiques strictement nécessaires à la tâche d'apprentissage s'alignent naturellement avec ce principe. Cependant, l'application pratique de ce concept nécessite une analyse rigoureuse des flux d'information spécifiques à chaque déploiement.

Le "droit à l'oubli" pose des défis techniques particuliers dans un contexte fédéré. Contrairement aux systèmes centralisés où les données peuvent être directement supprimées, l'influence des données d'un utilisateur spécifique est diffusée à travers les paramètres du modèle global de manière difficile à isoler. Des recherches récentes explorent des mécanismes de "désapprentissage" (*unlearning*) permettant de retirer efficacement l'influence des données spécifiques sans nécessiter un réentraînement complet du modèle.

Certification et Standards Émergents

Face aux incertitudes réglementaires, plusieurs initiatives de standardisation émergent pour établir des pratiques communes en matière de sécurité et de confidentialité pour l'Apprentissage Fédéré. Ces standards visent à fournir un cadre de référence tant pour les développeurs que pour les autorités de régulation.

L'ISO/IEC est en train de développer des extensions à ses standards existants sur la protection de la vie privée (série 27000) pour couvrir spécifiquement les paradigmes d'apprentissage distribué comme l'Apprentissage Fédéré. Ces extensions définiraient des exigences minimales en termes de protection contre les attaques par inférence et de robustesse face aux manipulations malveillantes.

Parallèlement, des consortiums industriels comme l'Open Federated Learning (OpenFL) travaillent à l'établissement de standards techniques spécifiques au domaine. Ces initiatives définissent des protocoles de communication sécurisés, des métriques standardisées pour l'évaluation des risques de fuite d'information, et des cadres d'audit permettant de vérifier la conformité des implémentations.

Li et al. soulignent l'importance de ces standards pour faciliter l'interopérabilité sécurisée entre différentes implémentations d'Apprentissage Fédéré. Sans ces références communes, chaque déploiement risque de développer des approches sécuritaires incompatibles, limitant ainsi les possibilités de collaboration à grande échelle qui constituent l'une des promesses fondamentales de cette technologie.

Considérations Éthiques et Sociétales

Équité Algorithmique et Biais dans un Contexte Fédéré

La question de l'équité algorithmique, déjà complexe dans les systèmes d'apprentissage traditionnels, prend une dimension supplémentaire dans le contexte fédéré. La nature décentralisée de l'apprentissage peut soit atténuer soit amplifier les biais présents dans les données, selon les mécanismes d'agrégation et les distributions des données entre participants.

D'une part, l'Apprentissage Fédéré offre l'opportunité d'inclure des populations plus diverses dans le processus d'apprentissage, potentiellement réduisant les biais liés à la sous-représentation de certains groupes. D'autre part, les mécanismes de sécurité eux-mêmes peuvent introduire des iniquités subtiles. Par exemple, les techniques de confidentialité différentielle peuvent avoir un impact disproportionné sur les groupes minoritaires dont les patterns distinctifs sont plus facilement masqués par l'ajout de bruit.

Li et al. décrivent comment certains mécanismes de défense contre les attaques par empoisonnement, particulièrement ceux basés sur la détection d'anomalies statistiques, risquent d'écarter légitimement les contributions de participants dont les données diffèrent significativement de la majorité. Ce phénomène peut renforcer insidieusement les biais existants en marginalisant algorithmiquement les perspectives minoritaires.

Des recherches récentes explorent des approches d'agrégation "équitable" qui garantissent une influence minimale à tous les groupes démographiques pertinents, indépendamment de leur taille relative dans la population des participants. Ces méthodes visent à réconcilier les impératifs de sécurité avec les considérations d'équité, reconnaissant que la protection contre les manipulations malveillantes ne doit pas se faire au détriment de la diversité des perspectives.

Accessibilité et Fracture Numérique

Les exigences de sécurité et de confidentialité de l'Apprentissage Fédéré soulèvent également des questions d'accessibilité et d'inclusion technologique. Les mécanismes de protection sophistiqués imposent souvent des contraintes computationnelles significatives qui peuvent exclure certains dispositifs ou utilisateurs.

Mothukuri et al. notent que les approches cryptographiques avancées comme le chiffrement homomorphe ou le calcul multi-parties sécurisé nécessitent des ressources considérables, potentiellement inaccessibles aux appareils d'entrée de gamme ou plus anciens. Cette situation risque d'exacerber la fracture numérique existante, où les avantages de la participation à l'apprentissage collaboratif seraient réservés aux possesseurs des dispositifs les plus récents et puissants.

Ces considérations sont particulièrement critiques dans des secteurs comme la santé mobile ou les services financiers dans les pays en développement, où l'Apprentissage Fédéré pourrait apporter des bénéfices significatifs précisément aux populations disposant d'un accès limité aux technologies avancées.

Des initiatives de recherche visent à développer des "mécanismes de sécurité inclusifs" qui offrent des garanties fondamentales même sur des dispositifs aux ressources limitées. Ces approches

explorent des compromis adaptifs où le niveau de protection s'ajuste aux capacités du dispositif, assurant une inclusion maximale tout en maintenant un seuil minimal de sécurité.

Gouvernance et Responsabilité Partagée

Modèles de Gouvernance pour les Fédérations Sécurisées

La nature distribuée de l'Apprentissage Fédéré nécessite des modèles de gouvernance innovants qui définissent clairement les responsabilités en matière de sécurité et de confidentialité. Contrairement aux systèmes centralisés où les responsabilités sont concentrées, les environnements fédérés impliquent un partage complexe des obligations entre multiples parties prenantes.

Li et al. proposent plusieurs modèles de gouvernance adaptés à différents contextes de déploiement. Dans le modèle "centralisé-responsable", une entité centrale (généralement le fournisseur de la plateforme fédérée) assume la responsabilité principale de la sécurité du système, définissant et imposant des politiques uniformes à tous les participants. Ce modèle simplifie la coordination mais crée une asymétrie de pouvoir potentiellement problématique.

À l'opposé, le modèle "décentralisé-démocratique" distribue les responsabilités de sécurité entre tous les participants, qui collectivement définissent et font évoluer les politiques par des mécanismes de consensus. Ce modèle plus égalitaire pose cependant des défis significatifs en termes de coordination et d'application effective des politiques.

Des approches hybrides émergent également, comme les modèles "fédération de fédérations" où des groupes de participants forment des sous-fédérations avec leurs propres politiques internes de sécurité, tout en adhérant à un ensemble minimal de standards communs pour l'interopérabilité. Cette structure offre un équilibre intéressant entre autonomie locale et cohérence globale.

Transparence et Auditabilité des Mécanismes de Sécurité

La transparence des mécanismes de sécurité constitue un principe fondamental pour établir la confiance dans les systèmes d'Apprentissage Fédéré. Pourtant, cette transparence doit être soigneusement calibrée pour ne pas compromettre l'efficacité même des protections mises en place.

Mothukuri et al. soulignent le paradoxe de la "sécurité par l'obscurité" dans ce contexte: d'une part, des mécanismes entièrement transparents peuvent être plus facilement contournés par des adversaires sophistiqués; d'autre part, des systèmes opaques risquent de ne pas inspirer confiance aux participants ou aux régulateurs.

Des approches innovantes tentent de résoudre ce dilemme par le concept d'"auditabilité sélective". Ces systèmes permettent à des auditeurs indépendants de vérifier certains aspects critiques des mécanismes de sécurité sans révéler publiquement tous les détails techniques qui pourraient être exploités par des attaquants. Des techniques cryptographiques avancées comme les preuves à divulgation nulle de connaissance permettent de démontrer mathématiquement certaines propriétés de sécurité sans révéler les mécanismes sous-jacents.

La standardisation des processus d'audit de sécurité spécifiques à l'Apprentissage Fédéré reste cependant un domaine émergent. Contrairement aux systèmes centralisés où des méthodologies

d'audit bien établies existent, l'évaluation de la sécurité des systèmes fédérés nécessite encore des cadres adaptés à leur nature distribuée et aux risques spécifiques qu'ils présentent.

8. Conclusion: Vers un Apprentissage Fédéré Sécurisé

Au terme de cette analyse approfondie des aspects sécuritaires et de confidentialité de l'Apprentissage Fédéré, plusieurs constats fondamentaux émergent qui définissent les enjeux présents et futurs de cette technologie prometteuse.

Équilibre des Forces et Faiblesses

L'Apprentissage Fédéré représente indéniablement une avancée significative dans la réconciliation entre l'exploitation collaborative des données et la protection de la vie privée. En maintenant les données brutes sur les dispositifs des utilisateurs ou au sein des organisations, cette approche élimine de facto certaines vulnérabilités traditionnelles liées à la centralisation des données sensibles. Cependant, comme l'ont démontré Mothukuri et al. ainsi que Li et al., cette décentralisation s'accompagne de nouvelles surfaces d'attaque et de vulnérabilités spécifiques.

La richesse des attaques documentées – empoisonnement, inférence, reconstruction, porte dérobée – témoigne d'une tension intrinsèque: le partage des mises à jour de modèles, nécessaire à l'apprentissage collaboratif, constitue inévitablement un canal d'information pouvant être exploité par des adversaires sophistiqués. Cette réalité fondamentale impose d'abandonner toute vision naïvement optimiste qui présenterait l'Apprentissage Fédéré comme une solution absolue aux problèmes de confidentialité.

L'analyse des mécanismes de défense révèle quant à elle un trilemme persistant entre confidentialité, performance et efficacité computationnelle. Les approches cryptographiques offrent des garanties théoriques fortes mais imposent des surcharges prohibitives; la confidentialité différentielle préserve la vie privée au prix d'une dégradation du modèle; les techniques d'agrégation robuste protègent contre l'empoisonnement mais risquent d'écarter des contributions légitimes mais atypiques. Ce trilemme ne représente pas un échec de l'ingénierie actuelle mais reflète des contraintes fondamentales probablement insurmontables dans leur totalité.

Perspectives d'Évolution Sécuritaire

Malgré ces défis, les directions de recherche identifiées offrent des perspectives encourageantes pour améliorer significativement l'équilibre entre sécurité, confidentialité et performance. L'approche contextuelle et adaptative de la sécurité, qui module les mécanismes de protection en fonction des risques spécifiques et des ressources disponibles, semble particulièrement prometteuse pour des déploiements pratiques.

L'intégration de couches de protection complémentaires – combiner par exemple la confidentialité différentielle locale avec l'agrégation sécurisée et des mécanismes de détection d'anomalies – offre une résilience accrue face à la diversité des menaces. Cette approche multi-niveaux, inspirée des principes de "défense en profondeur" établis en cybersécurité, paraît particulièrement adaptée à la complexité des systèmes fédérés.

Les études de cas présentées démontrent que des implémentations pratiques peuvent effectivement atteindre un équilibre viable entre protection et utilité, à condition d'être soigneusement calibrées pour leur contexte spécifique. La santé, la finance et les applications mobiles présentent des profils de risque distincts qui nécessitent des architectures de sécurité adaptées, plutôt qu'une approche universelle.

L'émergence de standards et de certifications spécifiques à l'Apprentissage Fédéré constituera probablement un catalyseur important pour son adoption à grande échelle. En établissant des références communes d'évaluation et des niveaux minimaux de protection attendus, ces standards faciliteront la comparaison objective des solutions et renforceront la confiance des utilisateurs comme des régulateurs.

Réflexions sur l'Avenir de la Confidentialité Distribuée

Au-delà des aspects purement techniques, l'avenir de l'Apprentissage Fédéré sécurisé dépendra largement de sa capacité à s'intégrer harmonieusement dans des écosystèmes socio-techniques complexes. Les modèles de gouvernance, les cadres réglementaires et les considérations éthiques joueront un rôle aussi déterminant que les avancées algorithmiques.

La réussite de cette intégration reposera sur un équilibre délicat entre plusieurs facteurs parfois contradictoires. D'une part, les mécanismes de sécurité doivent être suffisamment robustes pour résister à des adversaires sophistiqués et évolutifs. D'autre part, ces protections doivent rester accessibles et inclusives, évitant de créer une nouvelle fracture numérique entre ceux qui peuvent participer à l'apprentissage collaboratif et ceux qui en seraient exclus par manque de ressources computationnelles.

De même, la transparence des mécanismes de protection doit être calibrée avec soin: suffisante pour inspirer confiance aux participants et satisfaire les exigences réglementaires, mais sans compromettre l'efficacité même des protections en place. Les approches d'auditabilité sélective et de vérifiabilité formelle offrent des pistes prometteuses pour résoudre cette tension.

Enfin, il convient de reconnaître que l'Apprentissage Fédéré, malgré ses promesses, ne constituera jamais une solution universelle à tous les défis de confidentialité et de sécurité dans l'intelligence artificielle. Certains contextes aux exigences extrêmes de confidentialité ou confrontés à des adversaires particulièrement motivés et puissants pourront nécessiter des approches encore plus conservatrices, voire l'abstention pure et simple de certaines formes d'apprentissage collaboratif.

En définitive, l'Apprentissage Fédéré représente une évolution significative plutôt qu'une révolution absolue dans la réconciliation entre les bénéfices de l'intelligence artificielle collaborative et les impératifs de protection des données. Sa valeur réside précisément dans sa capacité à élargir l'espace des compromis viables, permettant dans de nombreux contextes de déplacer favorablement l'équilibre entre utilité et protection. Le chemin vers sa maturité impliquera une vigilance constante face aux vulnérabilités émergentes et une innovation continue dans les mécanismes de protection, guidées par une compréhension approfondie des contextes spécifiques de déploiement et des modèles de menace pertinents.## 7. Implications Réglementaires et Éthiques

Au-delà des aspects purement techniques, la sécurité et la confidentialité de l'Apprentissage Fédéré soulèvent des questions réglementaires et éthiques complexes qui influencent profondément son déploiement dans différents contextes. Cette section explore ces dimensions souvent négligées mais cruciales pour l'adoption responsable de cette technologie.

Cadres Réglementaires et Conformité

Conformité aux Réglementations de Protection des Données

L'Apprentissage Fédéré émerge dans un paysage réglementaire en pleine évolution concernant la protection des données personnelles. Si cette approche semble naturellement alignée avec l'esprit des législations comme le RGPD européen ou le CCPA californien, son implémentation pratique soulève néanmoins des questions d'interprétation juridique significatives.

Mothukuri et al. soulignent que malgré l'absence de centralisation directe des données brutes, certains aspects de l'Apprentissage Fédéré pourraient toujours tomber sous le coup de ces réglementations. Les mises à jour de modèles partagées pourraient, dans certaines circonstances, être considérées comme des "données personnelles" au sens juridique si elles permettent potentiellement l'identification d'individus via des attaques par inférence sophistiquées.

La notion de "minimisation des données", principe fondamental du RGPD, trouve une résonance particulière dans le contexte fédéré. Les implémentations qui extraient et partagent uniquement les caractéristiques strictement nécessaires à la tâche d'apprentissage s'alignent naturellement avec ce principe. Cependant, l'application pratique de ce concept nécessite une analyse rigoureuse des flux d'information spécifiques à chaque déploiement.

Le "droit à l'oubli" pose des défis techniques particuliers dans un contexte fédéré. Contrairement aux systèmes centralisés où les données peuvent être directement supprimées, l'influence des données d'un utilisateur spécifique est diffusée à travers les paramètres du modèle global de manière difficile à isoler. Des recherches récentes explorent des mécanismes de "désapprentissage" (*unlearning*) permettant de retirer efficacement l'influence des données spécifiques sans nécessiter un réentraînement complet du modèle.

Certification et Standards Émergents

Face aux incertitudes réglementaires, plusieurs initiatives de standardisation émergent pour établir des pratiques communes en matière de sécurité et de confidentialité pour l'Apprentissage Fédéré. Ces standards visent à fournir un cadre de référence tant pour les développeurs que pour les autorités de régulation.

L'ISO/IEC est en train de développer des extensions à ses standards existants sur la protection de la vie privée (série 27000) pour couvrir spécifiquement les paradigmes d'apprentissage distribué comme l'Apprentissage Fédéré. Ces extensions définiraient des exigences minimales en termes de protection contre les attaques par inférence et de robustesse face aux manipulations malveillantes.

Parallèlement, des consortiums industriels comme l'Open Federated Learning (OpenFL) travaillent à l'établissement de standards techniques spécifiques au domaine. Ces initiatives définissent des

protocoles de communication sécurisés, des métriques standardisées pour l'évaluation des risques de fuite d'information, et des cadres d'audit permettant de vérifier la conformité des implémentations.

Li et al. soulignent l'importance de ces standards pour faciliter l'interopérabilité sécurisée entre différentes implémentations d'Apprentissage Fédéré. Sans ces références communes, chaque déploiement risque de développer des approches sécuritaires incompatibles, limitant ainsi les possibilités de collaboration à grande échelle qui constituent l'une des promesses fondamentales de cette technologie.

Considérations Éthiques et Sociétales

Équité Algorithmique et Biais dans un Contexte Fédéré

La question de l'équité algorithmique, déjà complexe dans les systèmes d'apprentissage traditionnels, prend une dimension supplémentaire dans le contexte fédéré. La nature décentralisée de l'apprentissage peut soit atténuer soit amplifier les biais présents dans les données, selon les mécanismes d'agrégation et les distributions des données entre participants.

D'une part, l'Apprentissage Fédéré offre l'opportunité d'inclure des populations plus diverses dans le processus d'apprentissage, potentiellement réduisant les biais liés à la sous-représentation de certains groupes. D'autre part, les mécanismes de sécurité eux-mêmes peuvent introduire des iniquités subtiles. Par exemple, les techniques de confidentialité différentielle peuvent avoir un impact disproportionné sur les groupes minoritaires dont les patterns distinctifs sont plus facilement masqués par l'ajout de bruit.

Li et al. décrivent comment certains mécanismes de défense contre les attaques par empoisonnement, particulièrement ceux basés sur la détection d'anomalies statistiques, risquent d'écarter légitimement les contributions de participants dont les données diffèrent significativement de la majorité. Ce phénomène peut renforcer insidieusement les biais existants en marginalisant algorithmiquement les perspectives minoritaires.

Des recherches récentes explorent des approches d'agrégation "équitable" qui garantissent une influence minimale à tous les groupes démographiques pertinents, indépendamment de leur taille relative dans la population des participants. Ces méthodes visent à réconcilier les impératifs de sécurité avec les considérations d'équité, reconnaissant que la protection contre les manipulations malveillantes ne doit pas se faire au détriment de la diversité des perspectives.

Accessibilité et Fracture Numérique

Les exigences de sécurité et de confidentialité de l'Apprentissage Fédéré soulèvent également des questions d'accessibilité et d'inclusion technologique. Les mécanismes de protection sophistiqués imposent souvent des contraintes computationnelles significatives qui peuvent exclure certains dispositifs ou utilisateurs.

Mothukuri et al. notent que les approches cryptographiques avancées comme le chiffrement homomorphe ou le calcul multi-parties sécurisé nécessitent des ressources considérables,

potentiellement inaccessibles aux appareils d'entrée de gamme ou plus anciens. Cette situation risque d'exacerber la fracture numérique existante, où les avantages de la participation à l'apprentissage collaboratif seraient réservés aux possesseurs des dispositifs les plus récents et puissants.

Ces considérations sont particulièrement critiques dans des secteurs comme la santé mobile ou les services financiers dans les pays en développement, où l'Apprentissage Fédéré pourrait apporter des bénéfices significatifs précisément aux populations disposant d'un accès limité aux technologies avancées.

Des initiatives de recherche visent à développer des "mécanismes de sécurité inclusifs" qui offrent des garanties fondamentales même sur des dispositifs aux ressources limitées. Ces approches explorent des compromis adaptifs où le niveau de protection s'ajuste aux capacités du dispositif, assurant une inclusion maximale tout en maintenant un seuil minimal de sécurité.

Gouvernance et Responsabilité Partagée

Modèles de Gouvernance pour les Fédérations Sécurisées

La nature distribuée de l'Apprentissage Fédéré nécessite des modèles de gouvernance innovants qui définissent clairement les responsabilités en matière de sécurité et de confidentialité. Contrairement aux systèmes centralisés où les responsabilités sont concentrées, les environnements fédérés impliquent un partage complexe des obligations entre multiples parties prenantes.

Li et al. proposent plusieurs modèles de gouvernance adaptés à différents contextes de déploiement. Dans le modèle "centralisé-responsable", une entité centrale (généralement le fournisseur de la plateforme fédérée) assume la responsabilité principale de la sécurité du système, définissant et imposant des politiques uniformes à tous les participants. Ce modèle simplifie la coordination mais crée une asymétrie de pouvoir potentiellement problématique.

À l'opposé, le modèle "décentralisé-démocratique" distribue les responsabilités de sécurité entre tous les participants, qui collectivement définissent et font évoluer les politiques par des mécanismes de consensus. Ce modèle plus égalitaire pose cependant des défis significatifs en termes de coordination et d'application effective des politiques.

Des approches hybrides émergent également, comme les modèles "fédération de fédérations" où des groupes de participants forment des sous-fédérations avec leurs propres politiques internes de sécurité, tout en adhérant à un ensemble minimal de standards communs pour l'interopérabilité. Cette structure offre un équilibre intéressant entre autonomie locale et cohérence globale.

Transparence et Auditabilité des Mécanismes de Sécurité

La transparence des mécanismes de sécurité constitue un principe fondamental pour établir la confiance dans les systèmes d'Apprentissage Fédéré. Pourtant, cette transparence doit être soigneusement calibrée pour ne pas compromettre l'efficacité même des protections mises en place.

Mothukuri et al. soulignent le paradoxe de la "sécurité par l'obscurité" dans ce contexte: d'une part, des mécanismes entièrement transparents peuvent être plus facilement contournés par des

adversaires sophistiqués; d'autre part, des systèmes opaques risquent de ne pas inspirer confiance aux participants ou aux régulateurs.

Des approches innovantes tentent de résoudre ce dilemme par le concept d'"auditabilité sélective". Ces systèmes permettent à des auditeurs indépendants de vérifier certains aspects critiques des mécanismes de sécurité sans révéler publiquement tous les détails techniques qui pourraient être exploités par des attaquants. Des techniques cryptographiques avancées comme les preuves à divulgation nulle de connaissance permettent de démontrer mathématiquement certaines propriétés de sécurité sans révéler les mécanismes sous-jacents.

La standardisation des processus d'audit de sécurité spécifiques à l'Apprentissage Fédéré reste cependant un domaine émergent. Contrairement aux systèmes centralisés où des méthodologies d'audit bien établies existent, l'évaluation de la sécurité des systèmes fédérés nécessite encore des cadres adaptés à leur nature distribuée et aux risques spécifiques qu'ils présentent.# Rapport sur l'Apprentissage Fédéré : Sécurité, Confidentialité et Perspectives d'Avenir

9. Références

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115, 619-640. https://doi.org/10.1016/j.future.2020.10.007

Li, Q., Diao, Y., Chen, Q., & He, B. (2023). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 3347-3366. https://doi.org/10.1109/TKDE.2021.3124599

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *International Conference on Artificial Intelligence and Statistics*, 2938-2948.

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. *IEEE Symposium on Security and Privacy (SP)*, 739-753.

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175-1191.

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. *NIPS Workshop on Private Multi-Party Machine Learning*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning.

Foundations and Trends in Machine Learning, 14(1-2), 1-210.

Wang, H., Kaplan, Z., Niu, D., & Li, B. (2020). Optimizing federated learning on non-iid data with reinforcement learning. *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, 1698-1707.