UQAC
Université du Québec
à Chicoutimi

SÉCURISATION DE L'APPRENTISSAGE FÉDÉRÉ PAR FILTRAGE DYNAMIQUE BASÉ SUR LA RÉPUTATION

PAR ISSIAKA ISCHOLLA BABATOUNDE MAZU

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

L'apprentissage fédéré (FL) permet d'entraîner un modèle global à partir de données réparties chez de multiples clients, sans centraliser les données brutes. Cette décentralisation s'accompagne toutefois de défis majeurs : hétérogénéité statistique (non-IID), comportements opportunistes (*free-riding*) et attaques d'empoisonnement, qui dégradent la robustesse et l'équité du processus d'agrégation. Évaluer de manière fiable et juste la contribution de chaque client devient alors un enjeu central.

Nous proposons **RBFL** (*Reputation-Based Federated Learning*), une approche qui combine (i) une *utilité directionnelle* fondée sur la similarité cosinus des gradients, (ii) une *valorisation équitable* par approximation **TMC-Shapley** (Monte-Carlo tronqué, avec arrêt précoce) pour estimer la contribution marginale des clients, et (iii) une *réputation dynamique* (EMA) qui pondère au fil des rondes l'influence des participants et filtre les mises à jour déviantes. Les mises à jour locales sont ainsi agrégées selon un score de confiance, qui promeut les clients cohérents et atténue l'impact des contributeurs nuisibles.

L'évaluation empirique couvre des partitions IID, non-IID et *single-label* sur MNIST, Adult et FetalHealth, avec 100 clients (sélection aléatoire de 50% par ronde) et une attaque *label-flipping* déclenchée en cours d'apprentissage. RBFL est comparé à *FedAvg*, *Leave-One-Out*, la *Shapley exacte*, ainsi qu'à une approche récente basée sur SHAP pour la détection d'attaques. Les résultats montrent (1) une meilleure tolérance à l'empoisonnement, notamment en contexte non-IID extrême, (2) des évaluations de contribution fidèles tout en réduisant le coût par rapport à la Shapley exacte (accélération sensible grâce à TMC-Shapley), et (3) une amélioration de la précision globale par pondération réputationnelle.

Les expériences portent volontairement sur l'attaque *label-flipping* ; l'extension à des menaces plus avancées (Sybil, collusions, *backdoor*) et la validation à très grande échelle constituent des perspectives directes. Dans l'ensemble, RBFL offre une voie pragmatique pour des systèmes FL plus **sûrs**, **équitables** et **scalables** dans des domaines sensibles (santé, finance, industrie).

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

FL *Federated Learning* — apprentissage fédéré
RBFL *Reputation-Based Federated Learning*
HFL *Horizontal Federated Learning* — apprentissage fédéré horizontal
VFL *Vertical Federated Learning* — apprentissage fédéré vertical
FTL *Federated Transfer Learning*
Cross-device déploiement à grande échelle sur terminaux (smartphones, IoT)
Cross-silo déploiement entre organisations (banques, hôpitaux, entreprises)
IID *Independent and Identically Distributed*
Non-IID non indépendantes et identiquement distribuées
FedAvg *Federated Averaging*
FedProx *Federated Proximal*
FedNova *Federated Normalized Averaging*
SGD *Stochastic Gradient Descent*
ADAM *Adaptive Moment Estimation*
LR *Learning Rate* — taux d'apprentissage
DP *Differential Privacy* — vie privée différentielle
DP-SGD *Differentially Private SGD*
HE *Homomorphic Encryption* — chiffrement homomorphe
SMPC *Secure Multi-Party Computation* — calcul multipartite sécurisé
P2P *Peer-to-Peer* — pair à pair
SV *Shapley Value* — valeur de Shapley
TMC *Truncated Monte Carlo*
TMC-SV *TMC-Shapley Value*
LOO *Leave-One-Out*
SHAP *SHapley Additive exPlanations*
SVM *Support Vector Machine*
AUC *Area Under the Curve*
ROC *Receiver Operating Characteristic*
MNIST *Modified National Institute of Standards and Technology*
UCI *University of California, Irvine*
CTG *Cardiotocography* — cardiotocographie (FetalHealth)
RGPD Règlement Général sur la Protection des Données
GDPR *General Data Protection Regulation*
CPU *Central Processing Unit*

GPU *Graphics Processing Unit*
RAM *Random Access Memory*
IoT *Internet of Things* — Internet des objets
AICCSA *ACS/IEEE International Conference on Computer Systems and Applications*

DÉDICACE

À ma famille, source première de force et de motivation.

À mes parents, dont l'amour inconditionnel, les sacrifices silencieux et la sagesse ont tracé le chemin de ma persévérance et de ma réussite.

À mes frères et sœurs, pour leur inspiration, leur patience et la confiance qu'ils m'ont toujours témoignée, même dans les moments de doute.

À mes amis proches, qui par leur encouragement et leur bienveillance, ont su rendre ce parcours plus léger et plus riche humainement.

Enfin, à tous ceux qui, de près ou de loin, ont cru en moi et m'ont rappelé que chaque pas, aussi petit soit-il, mène vers la réalisation des plus grands rêves.

REMERCIEMENTS

Je souhaite exprimer ma profonde reconnaissance à toutes celles et ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire et à l'aboutissement de mon parcours académique.

Je tiens tout d'abord à remercier chaleureusement mon directeur de recherche, le professeur Fehmi Jaafar, pour son encadrement rigoureux, sa disponibilité constante et la confiance qu'il m'a accordée. Ses conseils avisés, sa vision scientifique et sa bienveillance ont été pour moi des repères précieux tout au long de ce travail.

Je tiens également à exprimer ma sincère gratitude à mon co-directeur de recherche, le professeur Hamdi Ben Abdessalem. Ses orientations méthodologiques, sa patience et ses remarques constructives ont contribué de manière essentielle à l'amélioration de ce mémoire. Son regard complémentaire a enrichi la qualité scientifique de ce projet et m'a permis de progresser de façon significative.

Mes remerciements vont également à Abdul Rehman, étudiant au doctorat, pour son aide précieuse, ses explications claires et son soutien technique lors des différentes étapes de mes expérimentations. Ses échanges stimulants et son esprit collaboratif ont eu un impact tangible sur l'avancement de mes recherches.

Je souhaite aussi exprimer ma gratitude à l'ensemble du corps professoral de l'Université du Québec à Chicoutimi pour la qualité des enseignements reçus, pour leur engagement envers la formation et pour l'inspiration qu'ils suscitent chez leurs étudiants.

Un remerciement particulier s'adresse à Mme Andréanne Riverin pour sa disponibilité et son soutien face à mes préoccupations administratives, ainsi qu'au professeur Bob-Antoine Jerry Ménélas, directeur du Département de Mathématiques et Informatique, pour son accompagnement et son dévouement.

Je n'oublie pas mes collègues et camarades de recherche, avec qui les discussions, les encouragements constants et le partage d'expériences ont enrichi mon parcours, tant sur le

plan académique que personnel.

Enfin, j'exprime toute ma gratitude à ma famille, véritable pilier de ce cheminement. Leur patience, leur compréhension et leur appui indéfectible m'ont permis d'avancer même dans les moments de doute et de fatigue. Je remercie aussi toutes celles et ceux, parfois anonymes, dont les gestes d'encouragement, les discussions ou les simples mots bienveillants ont nourri ma détermination et rendu ce parcours plus riche et plus humain.

AVANT-PROPOS

Ce mémoire s'inscrit dans la continuité des travaux de recherche visant à renforcer la sécurité et la fiabilité de l'apprentissage fédéré face aux menaces émergentes. Il propose une approche novatrice, baptisée **RBFL** (*Reputation-Based Federated Learning*), qui repose sur un mécanisme de filtrage adaptatif fondé sur la réputation afin de contrer efficacement les attaques par empoisonnement de données.

Dans un contexte où la quantité de données sensibles ne cesse de croître et où les systèmes distribués s'imposent progressivement comme un paradigme incontournable, les enjeux liés à la *sécurité*, à la *robustesse* et à l'*équité* deviennent cruciaux. Le présent travail ambitionne de répondre à ces défis en articulant une réflexion à la fois théorique, expérimentale et pragmatique, et en proposant une solution qui conjugue performance technique et applicabilité concrète.

Nous sommes également honorés de signaler que ce travail a été reconnu à l'échelle internationale, avec une publication acceptée à la conférence **IEEE AICCSA 2025 (International Conference on Computer Systems and Applications)**. Cette distinction atteste non seulement de la pertinence scientifique des problématiques abordées, mais aussi de l'intérêt croissant de la communauté académique et industrielle pour des solutions robustes et équitables en apprentissage fédéré.

Ce mémoire est ainsi le fruit d'une démarche collective et interdisciplinaire, portée par la volonté de contribuer à l'avancement des connaissances dans le domaine de l'intelligence artificielle sécurisée et de fournir des pistes concrètes pour des déploiements fiables dans des environnements distribués réels.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE DE LA RECHERCHE

À l'ère du numérique, l'intelligence artificielle (IA) irrigue des secteurs à haute criticité (santé, finance, télécommunications, industrie), où la valeur provient d'ensembles massifs de données hétérogènes. Or, la centralisation classique de ces données sur des serveurs distants se heurte à un double impératif : (i) la *conformité* aux cadres réglementaires de protection des données (p. ex. RGPD) et aux principes *privacy-by-design*, (ii) la *réduction des risques* de compromission (intrusions, fuites, mauvaise gouvernance des accès).

Dans ce contexte, l'**apprentissage fédéré** (*Federated Learning*, FL), introduit par McMahan *et al.* McMahan *et al.* (2016a), s'impose comme un paradigme de collaboration *décentralisée* : les données demeurent localement chez les clients (organisations, appareils), tandis qu'un modèle global est appris par échange de mises à jour (poids ou gradients) plutôt que de données brutes. Cette approche concilie souveraineté des données et efficacité statistique, et se décline tant en *cross-device* (très grand nombre d'appareils, connexions intermittentes) qu'en *cross-silo* (quelques entités institutionnelles, ressources stables, fortes contraintes de conformité).

Cependant, le FL introduit des défis structurels : (i) **hétérogénéité statistique** (*non-IID*) et opérationnelle (capacités de calcul, disponibilité, qualité des jeux locaux), (ii) **vulnérabilités** aux comportements malveillants (empoisonnement, *backdoor*, manipulation de gradients) et opportunistes (*free-riding*), (iii) **équité et incitations** : comment reconnaître, de façon *robuste* et *scalable*, la contribution effective de chaque participant au progrès du modèle global ? Ces

tensions rendent insuffisantes les agrégations naïves de type *FedAvg* dans des environnements réalistes et potentiellement adversariaux.

1.2 PROBLÉMATIQUE ET JUSTIFICATION

Le succès d'un système FL ne tient pas seulement au volume cumulé d'exemples locaux, mais à la *qualité*, la *fiabilité* et l'*équité* des mises à jour partagées. Or, plusieurs facteurs la compromettent :

- **Attaques d'empoisonnement** (p. ex. *label flipping*) qui biaisent la direction d'apprentissage et dégradent le modèle global ;
- **Comportements opportunistes** (*free-riding*) : certains clients cherchent à bénéficier du modèle sans contribuer réellement ;
- **Hétérogénéité extrême** : distributions non-IID ou *single-label*, entraînant des biais et des performances locales très variables ;
- **Contraintes de mesure côté serveur** : en pratique, l'agrégateur ne dispose pas toujours d'un jeu de validation central pour évaluer objectivement les contributions.

D'où la question centrale de ce mémoire :

*Comment assurer une évaluation **robuste**, **équitable** et **scalable** de la contribution des clients dans un environnement d'apprentissage fédéré soumis à l'hétérogénéité et aux menaces adversariales ?*

1.3 OBJECTIFS ET CONTRIBUTIONS DU MÉMOIRE

Pour répondre à cette problématique, nous proposons **RBFL** (*Reputation-Based Federated Learning*), une approche unifiée qui combine (i) une *utilité directionnelle* fondée sur la similarité cosinus de gradients, (ii) une *valorisation* par approximation *TMC-Shapley* pour

l'équité et la scalabilité, et (iii) une *réputation dynamique* multi-rondes pour promouvoir les contributeurs cohérents et atténuer l'influence des acteurs nuisibles.

OBJECTIFS

- **Mesurer sans données serveur** la pertinence immédiate des mises à jour grâce à une *utilité directionnelle* (similarité cosinus entre gradients locaux et direction globale), évitant la dépendance à un jeu de validation central.
- **Valoriser équitablement** la contribution de chaque client via une approximation *Truncated Monte Carlo Shapley* (TMC-Shapley) Li *et al.* (2024), fidèle aux principes de la valeur de Shapley Shapley (1953) mais praticable à grande échelle.
- **Assurer la robustesse** face à l'empoisonnement de type *label flipping* et aux *free-riders*, en pondérant/excluant progressivement les clients incohérents grâce à une *réputation* lissée (EMA) et un *seuil* τ ajustable.
- **Fournir un protocole de réglage** des hyperparamètres clés (seuil de réputation τ, facteur de lissage α, nombre de permutations m), incluant une procédure robuste *sans supervision* (médiane/MAD) et des recommandations pratiques.
- **Étudier la scalabilité et le coût** de bout en bout (communication, cosinus, réputation, TMC-Shapley) et dégager des lignes directrices pour des déploiements à plusieurs centaines de clients.

CONTRIBUTIONS

- **Formulation de RBFL** : nous proposons une architecture intégrée (Figure 4.1) combinant utilité directionnelle, estimation TMC-Shapley et réputation dynamique, ainsi qu'un algorithme complet d'*entraînement–évaluation–filtrage* (Algorithme 4.2).

— **Utilité « Cosine-Gradient » & TMC-Shapley** : nous proposons une définition d'une utilité *agnostique aux données* et l'intégration d'une approximation Shapley fondée sur *l'échantillonnage Monte Carlo* avec *arrêt précoce*, réduisant fortement le coût par rapport à la Shapley exacte tout en préservant l'ordre relatif des contributions.

— **Mécanisme de réputation multi-rondes** : nous proposons un schéma de mise à jour EMA des scores, intégrant un *seuil* τ pour atténuer ou écarter les clients non fiables, ainsi qu'un *filtrage réactif* lorsque la déviation persiste, afin de limiter l'impact des mises à jour adversariales ou aléatoires.

— **Évaluation empirique reproductible** : nous proposons une simulation centralisée (100 clients dont 50% sélectionnés par ronde), sur trois jeux de données (MNIST, Adult, FetalHealth), avec partitions IID / non-IID / *single-label*, et une **attaque *label flipping*** déclenchée tardivement. Nous comparons notre méthode à *FedAvg*, *Leave-One-Out* (LOO) et à la *Shapley exacte*. Les résultats montrent que RBFL *maintient la précision du modèle global* tout en identifiant et neutralisant efficacement les clients adversariaux, tandis que l'approximation TMC-Shapley procure simultanément un *gain de temps substantiel* (jusqu'à $\sim$45% par ronde).

— **Guidelines de scalabilité** : nous proposons une analyse des coûts ($\mathcal{O}(kd)$ pour communication/agrégation/cosinus ; TMC-Shapley modulable via m et arrêt précoce) et des recommandations pratiques (direction robuste, lissage de réputation, réglage de m, recours éventuel à quantification/sparsification).

— **Périmètre et limites assumées** : nous proposons de borner ce travail à l'étude de l'attaque *label flipping* et à une simulation centralisée (sans latences réseau), ce qui cadre l'interprétation des résultats et ouvre des perspectives pour des scénarios plus complexes (backdoor, attaques byzantines, Sybil/collusion, déploiements distribués réels).

1.4 ORGANISATION DU DOCUMENT

Ce mémoire est organisé conformément au plan suivant :

1. **Chapitre II — Notions de base** : rappels sur l'apprentissage automatique, limites du centralisé, principes et typologies de l'FL, algorithmes d'agrégation (FedAvg et variantes), menaces (non-IID, intégrité, confidentialité, disponibilité) et outils mathématiques mobilisés (similarité cosinus, valeur de Shapley et approximations, réputation).
2. **Chapitre III — Travaux connexes** : panorama critique des défenses et de l'évaluation de contribution en FL (approches statistiques, basées gradients, par valorisation, par réputation, hybrides), et positionnement de *RBFL*.
3. **Chapitre IV — Approche RBFL** : formalisation de la méthode (utilité directionnelle, TMC-Shapley, réputation), architecture et algorithme, hyperparamètres et choix de conception.
4. **Chapitre V — Étude expérimentale, discussion et validation** : protocole (environnement, jeux de données et partitions), *modèle de menace* (label flipping), baselines et métriques ; résultats, analyses (incluant *TMC-Shapley vs Shapley exacte*), équité contributive, coûts et scalabilité ; discussion intégrée (limites, validité) et recommandations.
5. **Conclusion générale** : synthèse des apports, limites et perspectives (extension du modèle de menace, passage à l'échelle plus large, déploiement réel).

Ce positionnement introduit la contribution principale de ce travail : une *méthode unifiée* qui renforce simultanément **robustesse**, **équité** et **scalabilité** dans des environnements fédérés réalistes, tout en restant *praticable* d'un point de vue computationnel.

CHAPITRE II

NOTIONS DE BASE

2.1 APPRENTISSAGE AUTOMATIQUE : DÉFINITIONS ET CATÉGORIES

L'apprentissage automatique (*Machine Learning*, ML) est une branche de l'intelligence artificielle qui vise à concevoir des systèmes capables d'apprendre automatiquement à partir de données, sans qu'il soit nécessaire de programmer explicitement les règles. L'objectif principal est de construire des modèles capables de **prédire, classifier ou découvrir des structures** dans les données historiques, et de généraliser ces connaissances à de nouvelles observations.

PRINCIPALES CATÉGORIES DE L'APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique se décline en plusieurs catégories, chacune adaptée à des types de problèmes spécifiques, comme illustré à la figure 2.1.

— **Apprentissage supervisé** : le modèle reçoit un ensemble de données étiquetées (x_i, y_i) et apprend à approximer une fonction f telle que $y_i \approx f(x_i)$. *Exemples :* classification d'images médicales, prévision de séries temporelles, détection de fraudes financières.

— **Apprentissage non supervisé** : le modèle travaille sur des données non étiquetées pour découvrir des structures cachées, comme des regroupements (*clustering*) ou des patterns sous-jacents. *Exemples :* segmentation de patients selon des profils de santé, analyse de comportements clients, réduction de dimensionnalité.

— **Apprentissage semi-supervisé** : combine des données partiellement étiquetées et non étiquetées pour améliorer la performance du modèle lorsqu'il est coûteux de produire des

labels. *Exemples :* classification de documents avec peu d'annotations, reconnaissance vocale.

— **Apprentissage par renforcement** : un agent apprend à prendre des décisions séquentielles en interagissant avec un environnement. Chaque action génère une récompense ou une pénalité, servant de signal pour optimiser la stratégie de décision. *Exemples :* robotique autonome, jeux vidéo, allocation dynamique de ressources dans le cloud.

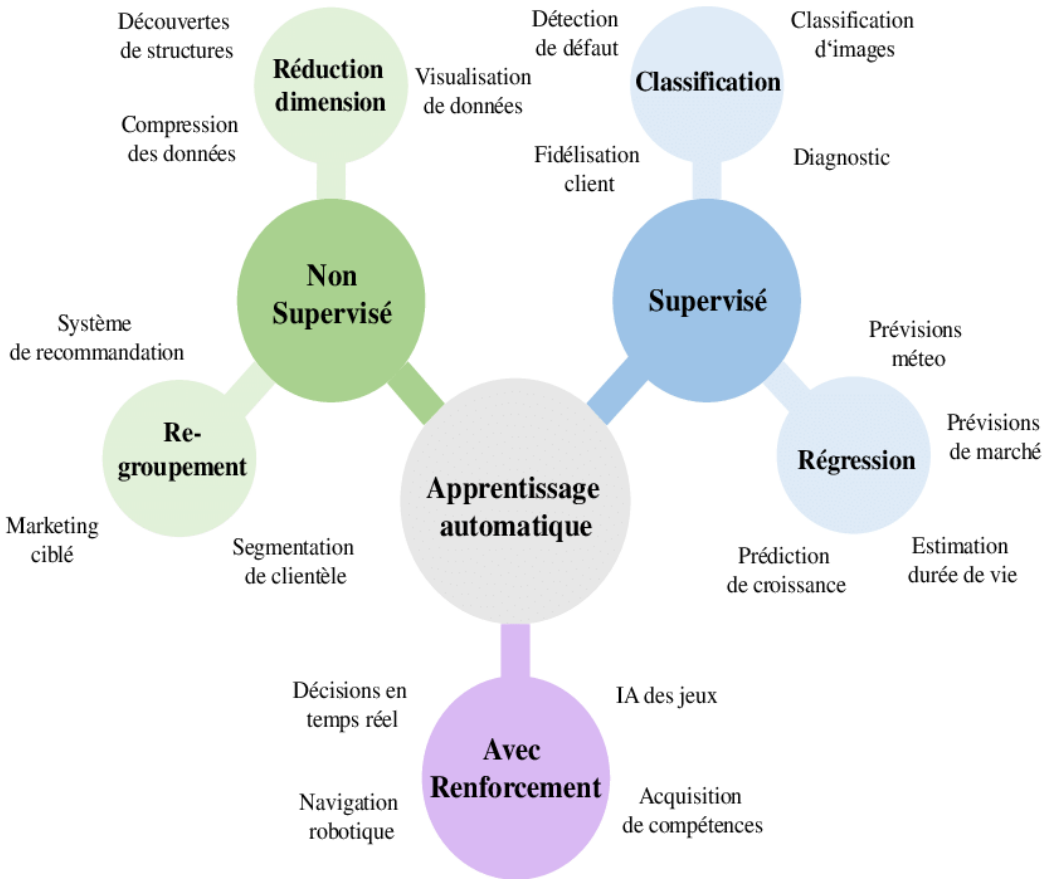

FIGURE 2.1 : Principales catégories d'apprentissage automatique et objectifs associés.
Riedel *et al.* (2024)

Cette section établit les bases nécessaires pour comprendre la distinction entre les modèles locaux et globaux dans les systèmes distribués, ainsi que les motivations qui conduisent

à l'apprentissage fédéré. La section suivante développera ces notions en détaillant les modèles locaux et globaux, les limites de l'apprentissage centralisé et les premières idées de décentralisation.

2.2 MODÈLES LOCAUX, MODÈLE GLOBAL ET LIMITES DU CENTRALISÉ

La Figure 2.2 illustre la relation entre les modèles locaux et le modèle global.

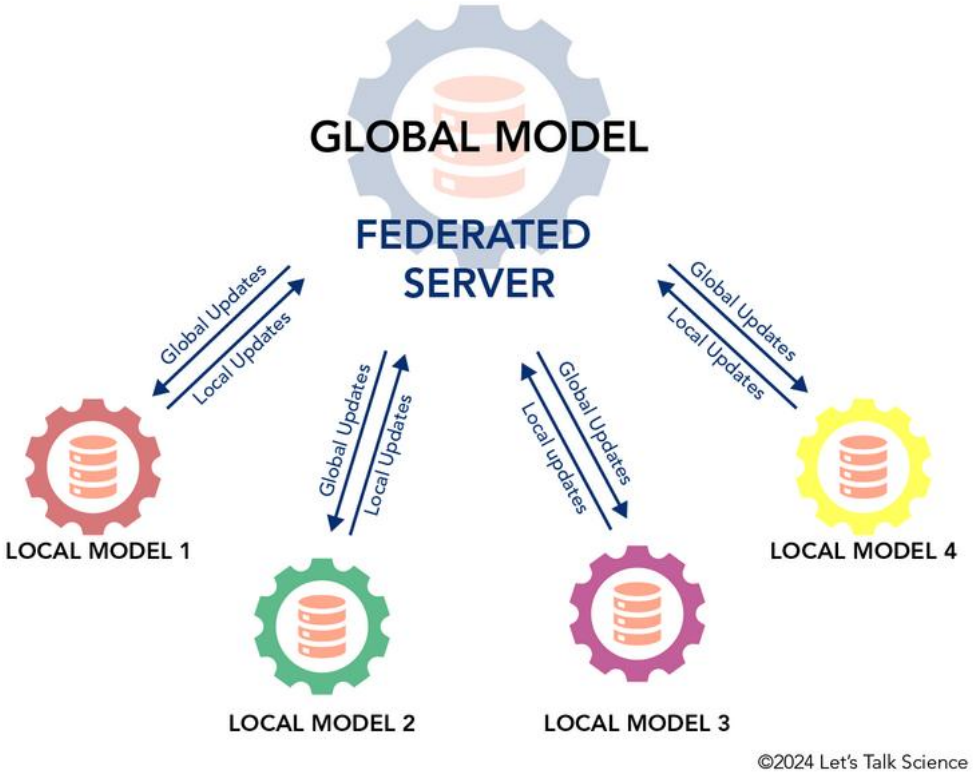

FIGURE 2.2 : Relation entre modèles locaux et modèle global dans un environnement distribué.
Sharma *et al.* (2023)

L'un des principes fondamentaux de l'apprentissage fédéré réside dans la distinction entre le **modèle local** et le **modèle global**. Chaque client entraîne son propre modèle sur ses données locales, appelé modèle local, tandis qu'un serveur central agrège l'ensemble de ces

contributions afin de produire un modèle global partagé. Ce mécanisme permet de trouver un équilibre subtil entre la spécificité locale, propre à chaque participant et la capacité de généralisation collective du modèle fédéré.

Le modèle local, entraîné exclusivement sur les données propres à chaque client, capture ainsi les spécificités individuelles et offre une personnalisation adaptée aux besoins locaux. Il garantit une confidentialité stricte, car aucune donnée brute n'est partagée, et sa rapidité d'entraînement constitue un atout dans de nombreux contextes. Toutefois, cette approche présente une capacité limitée de généralisation hors du domaine local, et il subsiste un manque de cohérence globale entre les différents clients.

À l'opposé, le modèle global naît de l'agrégation des modèles locaux. Il représente la synthèse collaborative de la connaissance du réseau, offrant une meilleure généralisation grâce à la diversité des données et une cohérence renforcée entre les participants. Cependant, cette stratégie dépend fortement de la qualité des mises à jour locales et nécessite une communication régulière et fiable avec le serveur central, ce qui peut poser des défis organisationnels et techniques.

Cette dualité structurante illustre l'avantage clé de l'apprentissage fédéré : permettre l'exploitation collective de données distribuées tout en respectant la confidentialité et la personnalisation, là où les approches centralisées échouent souvent à concilier ces exigences.

Le tableau 2.1 présente une comparaison entre les modèles locaux et le modèle global, en mettant en évidence leurs principales caractéristiques.

TABLEAU 2.1 : Comparaison entre modèles locaux et modèle global

Aspect	Modèle local	Modèle global
Source	Données d'un client	Agrégation des modèles locaux
Confidentialité	Très élevée	Élevée (pas de données brutes échangées)
Généralisation	Limitée	Forte
Complexité	Faible (côté client)	Élevée (côté serveur)
Communication	Nulle ou minimale	Essentielle
Adaptation	Personnalisée	Standardisée

2.2.1 LIMITES DE L'APPRENTISSAGE CENTRALISÉ

Avant l'émergence du FL, l'approche dominante consistait à rassembler toutes les données sur un serveur central pour entraîner un modèle global. Bien que simple à concevoir et à déployer, ce paradigme montre rapidement ses limites dans les environnements modernes.

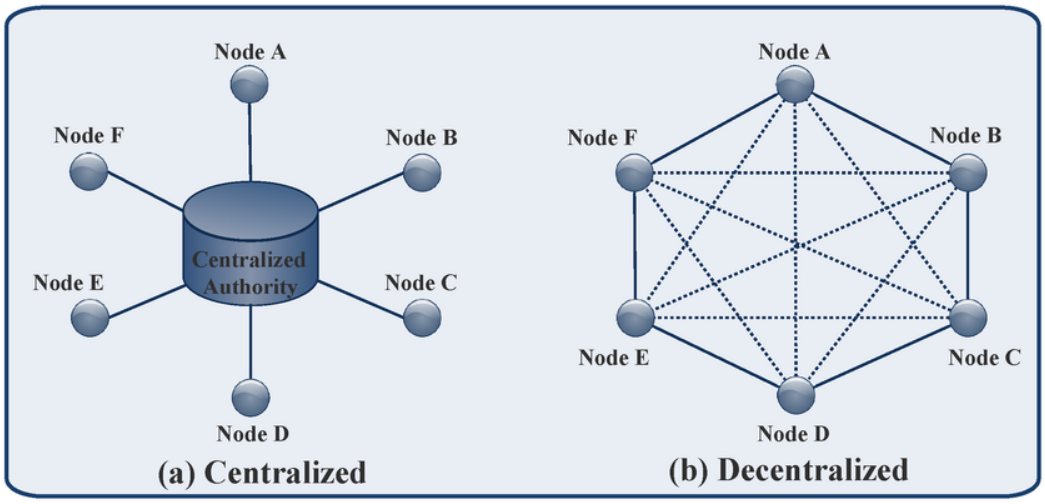

FIGURE 2.3 : Apprentissage centralisé et apprentissage distribué.

Problèmes de confidentialité : Le transfert massif des données brutes vers un point unique accroît le risque de divulgation d'informations sensibles et complexifie la conformité aux

régulations telles que le RGPD ou la HIPAA. Le serveur central devient par ailleurs un point de défaillance et une cible privilégiée, exposant l'ensemble du système à des attaques susceptibles d'entraîner des fuites à grande échelle.

Scalabilité et communication : À mesure que le nombre de clients augmente, les volumes de données à transporter rendent les coûts de communication et la latence prohibitifs. Le serveur central se transforme en goulot d'étranglement, ce qui dégrade les performances globales et alourdit l'infrastructure réseau nécessaire au maintien d'un débit acceptable.

Coûts computationnels : Concentrer l'entraînement sur une seule entité exige des capacités de calcul et de stockage importantes, difficiles à faire évoluer au rythme de la croissance des données et du nombre de clients. L'intégration de nouvelles sources devient coûteuse et lente, compromettant l'agilité du système.

Pour surmonter ces limitations, l'apprentissage distribué propose de **décentraliser l'entraînement**. Chaque client conserve ses données localement et entraîne un modèle sur son propre corpus, tandis que seules des informations dérivées (poids, gradients ou mises à jour compressées) sont partagées. Cette organisation préserve la confidentialité des données, réduit sensiblement les coûts de communication et permet au modèle global de bénéficier de la diversité statistique des jeux de données locaux sans nécessiter leur centralisation. La Figure 2.3 illustre clairement l'architecture de ces deux approches.

EXEMPLE CONCRET

Imaginons un réseau d'hôpitaux cherchant à élaborer un modèle prédictif pour le diagnostic de maladies chroniques. Dans une architecture centralisée classique, l'ensemble

des dossiers patients serait regroupé sur un serveur unique, soulevant aussitôt de sérieux enjeux de confidentialité et d'utilisation de la bande passante. Ce type de centralisation entre difficilement en résonance avec les exigences réglementaires, notamment celles qui entourent la protection des données médicales et la gestion sécurisée des flux d'information.

À l'inverse, une approche distribuée (en particulier fédérée) permet à chaque hôpital d'entraîner un modèle local sur ses propres données, sans transmettre de données brutes. Seules les mises à jour de paramètres sont partagées vers un serveur d'orchestration qui agrège ces contributions et construit un modèle global performant. Ce mécanisme garantit à la fois la confidentialité des dossiers patients et la maîtrise de la charge de communication. Il illustre l'intérêt fondamental de l'apprentissage distribué comme étape préparatoire à l'apprentissage fédéré, approfondi dans la section suivante.

2.3 L'APPRENTISSAGE FÉDÉRÉ (FEDERATED LEARNING, FL)

2.3.1 DÉFINITION ET PRINCIPES FONDAMENTAUX

L'apprentissage fédéré (FL) est un paradigme d'apprentissage collaboratif et décentralisé. Il permet à plusieurs clients (appareils, serveurs ou organisations) de participer à l'entraînement d'un modèle global tout en conservant leurs données localement Kairouz *et al.* (2021); Yang *et al.* (2019).

Contrairement à l'apprentissage centralisé, les données ne quittent jamais les nœuds locaux : seules les mises à jour du modèle (poids ou gradients) sont transmises au serveur central pour agrégation.

Processus cyclique du FL (3 étapes) Le FL se déroule typiquement en trois phases répétées à chaque ronde t :

1 — **Model initialization** : le serveur initialise le modèle global w_0 (aléatoire ou pré-entraîné), sélectionne un sous-ensemble de clients $\mathscr{P}_t$ et *publie* w_t vers ces clients.

2 — **Local model training & upload** : chaque client $i \in \mathscr{P}_t$ entraîne localement w_t sur ses données D_i pendant E époques, obtient $w_t^{(i)}$ et *uploade* au serveur soit les poids $w_t^{(i)}$, soit la mise à jour $\Delta w_t^{(i)} = w_t^{(i)} - w_t$ (ou les gradients correspondants).

3 — **Global model aggregation & broadcast** : le serveur agrège les contributions (p. ex. FedAvg) pour mettre à jour

$$w_{t+1} = \sum_{i \in \mathscr{P}_t} \frac{n_i}{\sum_{j \in \mathscr{P}_t} n_j} w_t^{(i)} \quad \text{(ou équivalent en } \Delta w\text{)},$$

puis *rediffuse* w_{t+1} aux clients pour la ronde suivante.

Ce cycle se répète jusqu'au critère d'arrêt (convergence, budget de rondes, métrique cible). Dans **RBFL**, l'étape 3 est précédée par le calcul des similarités cosinus et la mise à jour des scores de réputation afin de pondérer (ou filtrer) les mises à jour avant agrégation.

2.3.2 TYPOLOGIES DU FL

GRANDES FAMILLES DE FL

Le FL se décline en plusieurs variantes selon la structure des données comme le présente le tableau 2.2 :

— **FL horizontal (HFL)** : les clients partagent le même espace de variables mais possèdent des exemples différents (ex. : plusieurs hôpitaux ayant les mêmes attributs médicaux mais sur des patients distincts).

— **FL vertical (VFL)** : les clients possèdent des variables différentes sur un même ensemble d'individus (ex. : une banque et une assurance partageant des informations complémentaires sur les mêmes clients).

— **FL transféré (FTL)** : les clients n'ont ni les mêmes exemples ni les mêmes variables, mais cherchent à transférer des représentations utiles pour l'entraînement collaboratif.

TABLEAU 2.2 : Résumé des grandes familles de FL

Type	Variables	Exemples	Cas typiques
HFL	Identiques	Différents	Réseaux hospitaliers, applications mobiles
VFL	Différentes	Identiques/partiels	Banque vs assurance
FTL	Différentes	Différents	Partenariats intersectoriels

AVANTAGES PRINCIPAUX DU FL

Au-delà de sa définition et de son principe général, l'apprentissage fédéré présente un ensemble d'atouts qui expliquent son adoption croissante dans des domaines sensibles comme la santé, la finance ou l'Internet des objets. En supprimant la nécessité de centraliser les données, il concilie efficacité algorithmique et respect des contraintes légales ou éthiques liées à la confidentialité. De plus, sa capacité à fédérer un grand nombre de participants hétérogènes en fait une approche prometteuse pour des environnements réels caractérisés par la diversité des ressources et la fragmentation des données. Les principaux avantages du FL sont résumés dans le tableau 2.3.

TABLEAU 2.3 : Principaux avantages de l'apprentissage fédéré

Avantage	Description
Confidentialité	Les données restent locales, limitant les risques de fuite et respectant les régulations (RGPD, HIPAA).
Efficacité de communication	Seules les mises à jour des modèles sont transmises, réduisant la charge réseau.
Personnalisation	Possibilité d'adapter le modèle global aux particularités locales de chaque client.
Scalabilité	Peut intégrer un grand nombre de clients sans nécessiter la centralisation des données.
Résilience	Même si certains clients sont indisponibles, l'entraînement global peut continuer.

DÉFIS ET LIMITES DU FL

Malgré ses avantages, l'apprentissage fédéré se heurte à plusieurs difficultés majeures. D'abord, la **distribution non identique des données** entre clients peut ralentir la convergence, induire des biais et compliquer l'agrégation efficace des mises à jour. Ensuite, l'**hétérogénéité des clients** (capacités de calcul et de mémoire, disponibilité irrégulière, qualité des données) génère des « stragglers » et des déséquilibres de contribution qui nuisent à la stabilité de l'entraînement. Sur le plan de la **sécurité et de l'intégrité**, la présence potentielle de clients malveillants expose le système à des attaques d'empoisonnement ou de rétro-ingénierie, exigeant des mécanismes de détection et de robustesse dédiés. Enfin, la **communication et la synchronisation** restent des contraintes fortes : la dépendance au réseau, les coupures, la latence et la coordination inter-clients impactent directement le débit d'apprentissage et la qualité du modèle global.

Architecture de référence. Le schéma classique s'articule autour d'un *serveur central* qui orchestre les rounds, diffuse le modèle global, reçoit les mises à jour locales (poids ou gradients) et réalise l'agrégation pour produire une nouvelle version du modèle. Les *clients*

distribués conservent leurs données en local et entraînent un modèle sur leur corpus propre avant de transmettre uniquement des informations dérivées au serveur, préservant ainsi la confidentialité tout en contribuant à l'amélioration du modèle global.

ENTRAÎNEMENT PAR ITÉRATIONS LOCALES

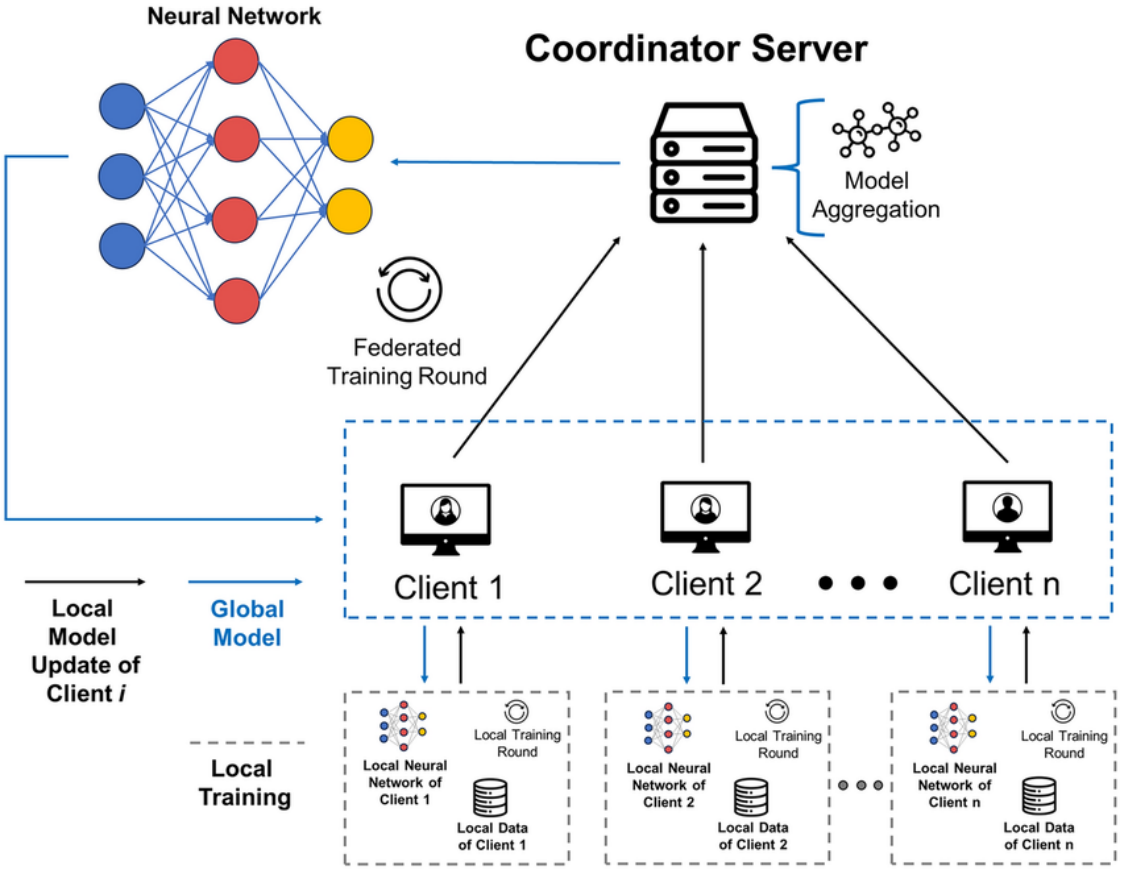

FIGURE 2.4 : Architecture détaillée d'un système d'apprentissage fédéré.
Riedel *et al.* (2024)

Le processus d'apprentissage fédéré se déroule sur plusieurs *rounds*, selon les étapes suivantes, comme illustré dans la Figure 2.4 :

1. Chaque client effectue E itérations de descente de gradient sur son modèle local.
2. Les paramètres locaux sont transmis au serveur.
3. Le serveur agrège les paramètres pour former le modèle global.
4. Le modèle global est renvoyé aux clients pour le prochain round.

2.3.3 EXEMPLE CONCRET : FL EN SANTÉ

Considérons un réseau de trois hôpitaux, comme l'illustre la Figure 2.5, souhaitant entraîner un modèle de détection précoce du diabète.

— Chaque hôpital conserve ses dossiers médicaux sur ses serveurs locaux.
— Chaque hôpital entraîne un modèle local et transmet uniquement les gradients au serveur central.
— Le serveur agrège les mises à jour pour obtenir un modèle global plus performant que chaque modèle local individuel.

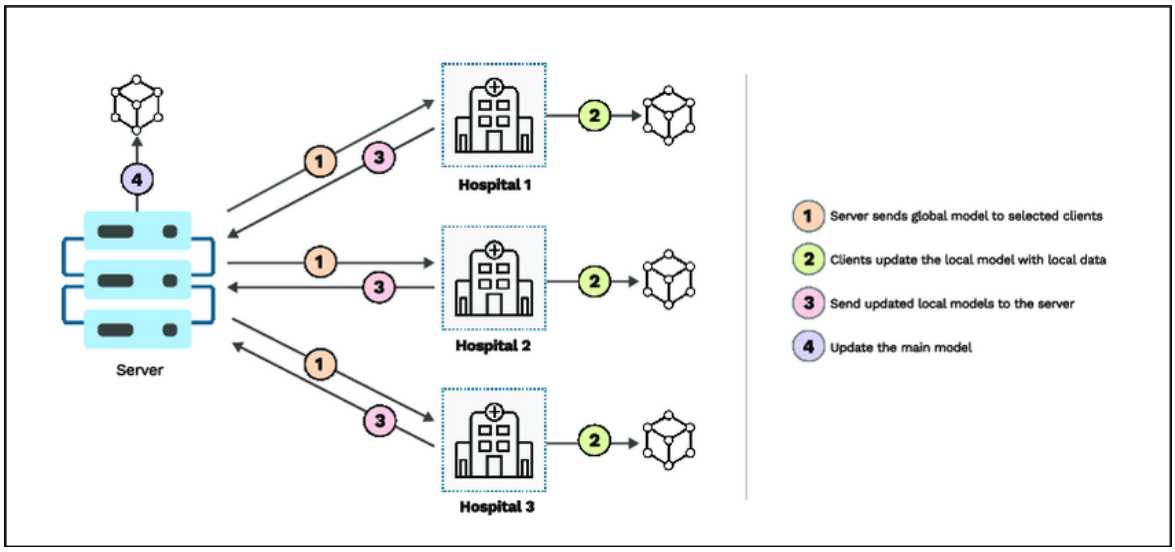

FIGURE 2.5 : Exemple de mise en œuvre de FL pour la santé.
Joshi *et al.* (2022)

2.3.4 FL VS APPRENTISSAGE CENTRALISÉ

Cette sous-section propose une analyse comparative entre l'apprentissage centralisé et l'apprentissage fédéré, dans le but de mieux comprendre leurs différences conceptuelles, comme le présente le tableau 2.4.

TABLEAU 2.4 : Comparaison entre apprentissage centralisé et FL

Critère	**Centralisé**	**Fédéré**
Données	Centralisées sur un serveur	Conservées localement
Communication	Transfert massif de données	Transmission uniquement des gradients
Confidentialité	Faible	Élevée
Scalabilité	Limitée	Très élevée
Adaptation locale	Nulle	Possible
Robustesse aux pannes	Faible	Plus élevée

2.4 ALGORITHMES D'AGRÉGATION EN FL

L'agrégation des modèles locaux est un élément central de l'apprentissage fédéré. Elle permet de combiner efficacement les modèles entraînés localement afin d'obtenir un modèle global cohérent et performant.

2.4.1 FEDAVG (FEDERATED AVERAGING)

FedAvg est l'algorithme de référence pour l'agrégation en FL, proposé par McMahan et al. McMahan *et al.* (2016a). Il repose sur une combinaison pondérée des paramètres locaux des clients. Comme l'illustre très bien la Figure 2.6.

FORMULATION MATHÉMATIQUE

Soit :

— K : nombre de clients participants,

— w_k^t : vecteur des paramètres du modèle local du client k après le round t,

— n_k : nombre d'exemples du client k,

— $n = \sum_{k=1}^{K} n_k$: nombre total d'exemples sur tous les clients.

Alors, le modèle global après le round $t+1$ est :

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^t \tag{2.1}$$

Chaque client effectue E itérations de descente de gradient locale avant de transmettre ses paramètres au serveur.

ILLUSTRATION

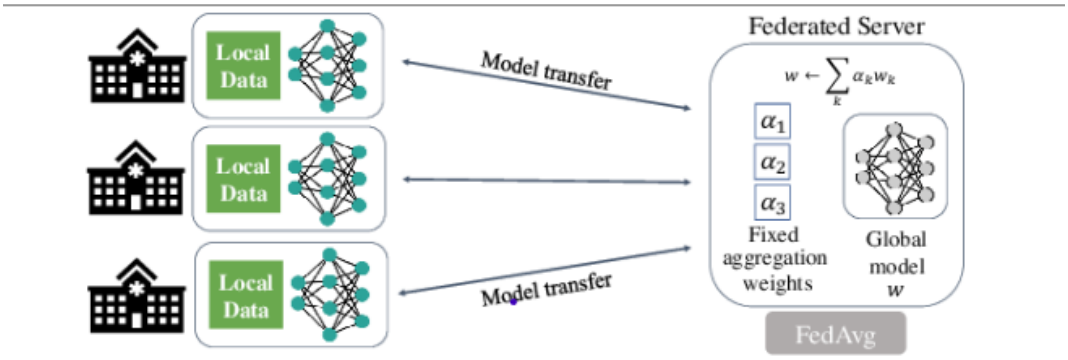

FIGURE 2.6 : Schéma illustrant l'agrégation FedAvg en apprentissage fédéré.
Xia *et al.* (2021)

AVANTAGES ET LIMITES DE FEDAVG

Le tableau 2.5 illustre les principaux avantages et limites de FedAvg.

TABLEAU 2.5 : Avantages et limites de FedAvg

	Avantages	**Limites**
Simplicité	Facile à implémenter et comprendre	Sensible aux données non-IID
Efficacité	Réduit le volume de communication	Peut diverger avec clients hétérogènes
Robustesse	Fonctionne avec différents nombres de clients	Vulnérable aux clients malveillants

2.4.2 VARIANTES DE FEDAVG

Plusieurs variantes ont été proposées pour améliorer la robustesse et la convergence :

— **FedProx** Zhu *et al.* (2024) : ajoute un terme de régularisation proximal pour limiter la divergence des modèles locaux sur des clients hétérogènes. La fonction de perte locale devient :

$$L_k(w) = L_k^{\text{local}}(w) + \frac{\mu}{2}\|w - w^t\|^2 \tag{2.2}$$

où μ est le coefficient de régularisation.

— **FedSGD** : agrégation basée sur la moyenne des gradients locaux plutôt que sur les poids du modèle. Adaptée aux environnements très distribués avec peu d'itérations locales.

— **Agrégation robuste** : inclut des techniques comme la *moyenne tronquée*, la *médiane* ou l'agrégation pondérée par la réputation des clients pour limiter l'impact des participants malveillants.

2.4.3 DISCUSSION

Les algorithmes présentés mettent en lumière l'importance centrale des mécanismes d'agrégation dans la robustesse et la performance de l'apprentissage fédéré. Bien que FedAvg constitue la pierre angulaire de nombreuses approches, il demeure nécessaire d'apporter des ajustements pour mieux gérer la forte hétérogénéité des données et des participants. Cette gestion est primordiale afin de limiter l'impact potentiellement déstabilisateur des clients malveillants, qui peuvent compromettre l'intégrité du modèle global. Par ailleurs, améliorer la convergence de l'algorithme demeure un enjeu crucial, en particulier dans les environnements très distribués où les ressources et la disponibilité des clients varient significativement.

La maîtrise fine de ces mécanismes d'agrégation est ainsi un préalable incontournable à toute analyse des vulnérabilités spécifiques à l'apprentissage fédéré et à l'élaboration des stratégies de défense adaptées. Ces aspects seront développés dans le chapitre suivant, qui se penchera plus précisément sur les risques liés à la sécurité et la résilience du système.

2.5 VULNÉRABILITÉS ET MENACES DANS L'APPRENTISSAGE FÉDÉRÉ

Malgré ses avantages en matière de confidentialité et de collaboration distribuée, l'apprentissage fédéré (FL) présente plusieurs vulnérabilités qui peuvent compromettre la performance, la robustesse et la sécurité du modèle global. Ces menaces peuvent être classées en quatre grandes catégories : données non-IID et biais statistique, attaques contre l'intégrité, attaques contre la confidentialité et attaques contre la disponibilité.

2.5.1 DONNÉES NON-IID ET HÉTÉROGÉNÉITÉ STATISTIQUE

Dans de nombreux scénarios FL, les données locales des clients ne sont pas *independent and identically distributed* (non-IID). Cette hétérogénéité peut se manifester par des déséquilibres de classes, des distributions spécifiques à certains clients ou des volumes de données variables.

— **Impact sur la convergence** : les modèles locaux divergent dans des directions différentes, ce qui peut ralentir la convergence du modèle global, comme l'illustre spécifiquement la Figure 2.7.

— **Biais statistique** : certaines classes ou populations peuvent être surreprésentées dans le modèle global, affectant la performance et l'équité.

— **Difficulté d'évaluation des contributions** : mesurer la contribution individuelle de chaque client devient complexe lorsque les données ne sont pas homogènes.

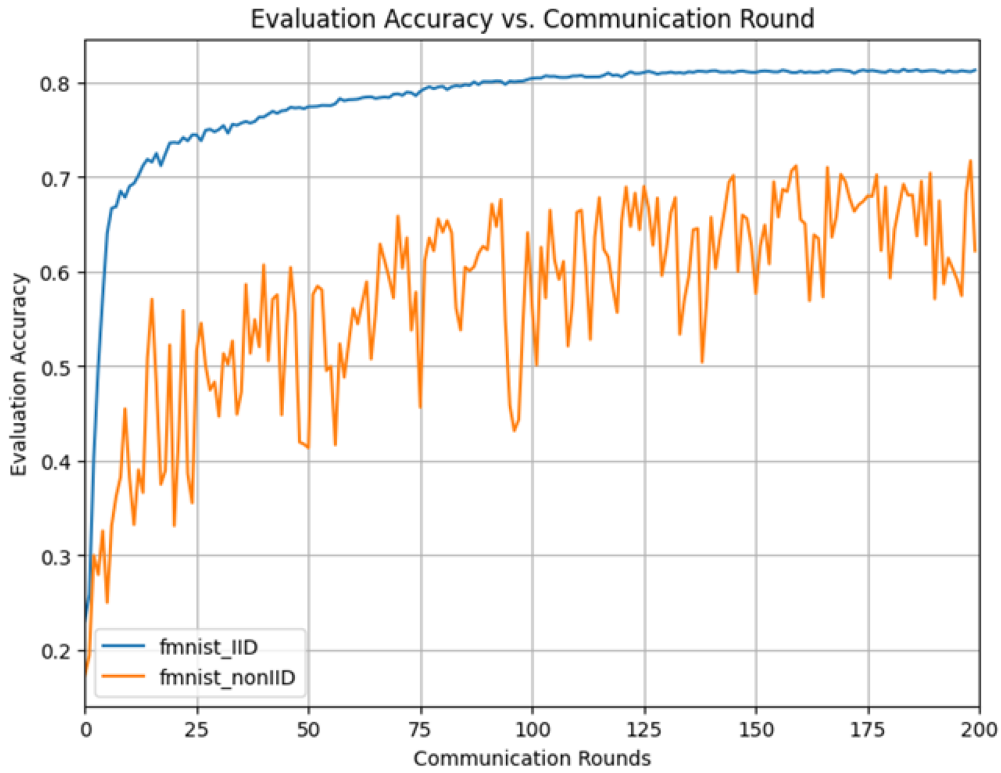

FIGURE 2.7 : Impact des données non-IID sur la convergence globale.
Efthymiadis *et al.* (2024)

2.5.2 ATTAQUES VISANT L'INTÉGRITÉ DU MODÈLE

Certains clients peuvent adopter un comportement malveillant dans le but de compromettre le modèle global. Le tableau 2.6 présente les formes d'attaques d'intégrité les plus couramment étudiées dans l'apprentissage fédéré :

— **Label-flipping** : modification intentionnelle des étiquettes du jeu de données local afin d'induire un biais dans l'apprentissage. Par exemple, inverser les classes « malade » et « sain » dans un modèle médical.

— **Attaques par *backdoor*** : insertion de motifs spécifiques (*triggers*) dans certaines entrées dans le but de forcer le modèle à produire des prédictions manipulées sur des cas particuliers.

TABLEAU 2.6 : Exemples d'attaques d'intégrité dans l'apprentissage fédéré

Type d'attaque	Principe	Impact sur le modèle
Label-flipping	Modification intentionnelle des étiquettes locales	Introduction d'un biais ciblé et dégradation des performances globales
Backdoor	Injection de déclencheurs (triggers) spécifiques dans les données locales	Activation d'un comportement malveillant lorsque des entrées particulières sont rencontrées
Poisoning numérique	Ajout de bruit ou envoi de gradients falsifiés	Perturbation de la convergence et réduction de l'efficacité globale du modèle

2.5.3 MENACES CONTRE LA CONFIDENTIALITÉ

Bien que les données demeurent stockées localement chez les clients, certaines attaques peuvent exploiter les gradients ou les modèles partagés afin de compromettre la confidentialité :

— **Attaques par inférence** : extraction d'informations sensibles à partir des gradients ou des mises à jour échangées. Un exemple parfaitement illustré à la figure 2.8.

— **Attaques par reconstruction** : reconstitution partielle, voire complète, des données locales d'un client à partir des informations partagées.

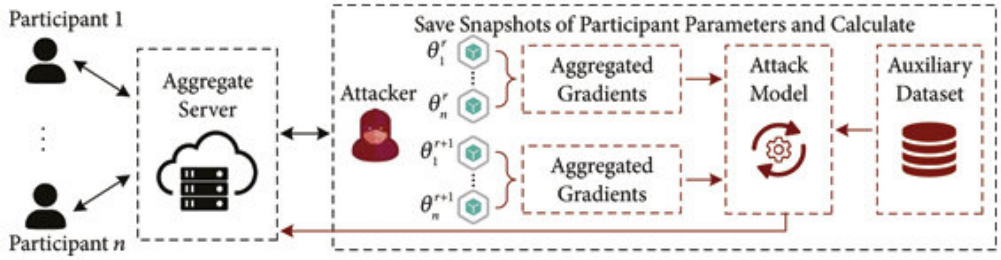

FIGURE 2.8 : Example of inference attacks in FL. The attacker saves the snapshots of the aggregated model parameters in each round and performs inference attacks by employing the difference between the continuous snapshots.
Zhang *et al.* (2022)

2.5.4 MENACES CONTRE LA DISPONIBILITÉ

Certaines attaques visent à perturber la disponibilité ou à réduire l'efficacité globale du système fédéré :

— **Attaques par déni de service (DoS)** : submersion du serveur central par des mises à jour corrompues ou massives, entraînant un blocage ou un ralentissement du processus d'agrégation.

— **Free-riders** : clients qui minimisent, voire évitent, leur participation au calcul local tout en cherchant à bénéficier du modèle global, ce qui compromet l'équité et l'efficacité du système.

Le Tableau 2.7 présente un résumé des principales catégories de vulnérabilités dans l'apprentissage fédéré.

TABLEAU 2.7 : Résumé des catégories de vulnérabilités dans FL

Catégorie	**Exemple**	**Conséquence**
Non-IID	Distributions locales différentes	Divergence, biais, difficulté d'évaluation
Intégrité	Label-flipping, backdoor	Dégradation performance, comportements malveillants
Confidentialité	Inference, reconstruction	Fuite d'informations sensibles
Disponibilité	DoS, free-riders	Ralentissement ou blocage de l'apprentissage

Cette section met en évidence l'importance d'intégrer des mécanismes de détection, de filtrage et d'évaluation des contributions pour assurer la robustesse et l'équité du FL.

2.6 NOTIONS MATHÉMATIQUES ET OUTILS FONDAMENTAUX

Pour formaliser des mécanismes robustes d'évaluation des clients dans l'apprentissage fédéré, il est essentiel de maîtriser certaines notions mathématiques et concepts clés. Ces outils servent à mesurer la similarité des mises à jour locales, quantifier la contribution individuelle et établir des mécanismes de réputation fiables.

2.6.1 SIMILARITÉ COSINUS

La similarité cosinus mesure la proximité directionnelle entre deux vecteurs, comme l'illustre la Figure 2.9. En FL, elle est utilisée pour comparer les gradients locaux des clients avec ceux du modèle global.

Définition 2.6.1. *Soient deux vecteurs* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. *La similarité cosinus est définie par :*

$$sim_{\cos}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} v_i^2}} \tag{2.3}$$

Interprétation :

- Une valeur proche de 1 indique que les gradients sont fortement alignés, traduisant une contribution cohérente et positive.
- Une valeur proche de 0 reflète une quasi-orthogonalité des gradients, suggérant une contribution neutre ou négligeable.
- Une valeur proche de -1 révèle une opposition marquée, pouvant signaler un comportement anormal ou malveillant.

Exemple en FL : Si **u** correspond au gradient d'un client et **v** au gradient global, une similarité de 0,95 traduit un alignement fort avec la direction d'optimisation collective. À

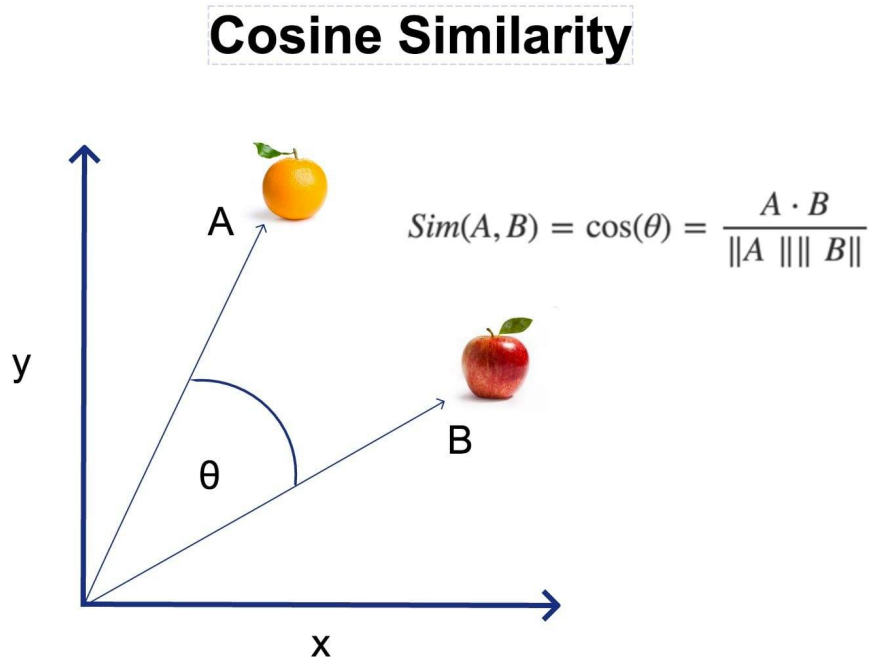

FIGURE 2.9 : Illustration géométrique de la similarité cosinus entre deux vecteurs dans un espace multidimensionnel.
Trend (2023)

l'inverse, une valeur de -0,8 peut être interprétée comme une tentative de perturbation du processus d'apprentissage.

2.6.2 VALEUR DE SHAPLEY

La valeur de Shapley, issue de la théorie des jeux coopératifs, fournit une mesure équitable de la contribution marginale d'un participant au sein d'une coalition. Elle permet de répartir de manière juste les gains générés collectivement en tenant compte de l'impact individuel de chaque joueur.

Définition 2.6.2. *Soit un jeu coopératif* (N, v)*, où* N *désigne l'ensemble des joueurs et* v : $2^N \to \mathbb{R}$ *une fonction de valeur attribuant un gain à chaque sous-ensemble de joueurs. La*

valeur de Shapley associée à un joueur $i \in N$ est définie par :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big[v(S \cup \{i\}) - v(S) \Big] \tag{2.4}$$

Application en apprentissage fédéré : Dans ce contexte, chaque client peut être vu comme un joueur, et la fonction de valeur $v(S)$ correspond à la performance du modèle global entraîné avec la coalition de clients S. La valeur de Shapley $\phi_i(v)$ représente alors la contribution marginale moyenne du client i, calculée sur l'ensemble des permutations possibles des participants.

TABLEAU 2.8 : Exemple simplifié de calcul de la valeur de Shapley pour trois clients

Client	Performance ajoutée (moyenne sur permutations)	Valeur de Shapley
1	0.05, 0.08, 0.06	0.063
2	0.02, 0.07, 0.03	0.040
3	0.01, 0.05, 0.02	0.027

Le Tableau 2.8 présente un exemple simplifié de calcul de la valeur de Shapley pour trois clients, illustrant la performance ajoutée moyenne sur les permutations et la valeur finale attribuée à chaque contributeur.

2.6.3 APPROXIMATIONS DE LA VALEUR DE SHAPLEY EN FL

Le calcul exact de la valeur de Shapley est de complexité exponentielle en fonction du nombre de clients $|N|$. Afin de rendre son utilisation praticable dans un contexte fédéré, plusieurs méthodes d'approximation ont été proposées.

TMC-SHAPLEY (TRUNCATED MONTE CARLO SHAPLEY)

— Repose sur un échantillonnage aléatoire d'un sous-ensemble de permutations de clients.

— Retient en priorité les permutations les plus représentatives pour accélérer le calcul.

— Réduit considérablement la complexité computationnelle tout en préservant une estimation suffisamment précise de la contribution.

GTG-SHAPLEY (GRADIENT TRUNCATION GUIDED SHAPLEY)

— Adaptation spécifique au contexte de l'apprentissage fédéré.

— Tronque les gradients extrêmes afin de limiter l'impact de mises à jour malveillantes.

— Combine les avantages de TMC-Shapley et des méthodes robustes basées sur les gradients, fournissant ainsi une estimation plus fiable.

2.6.4 RÉPUTATION ET FILTRAGE DYNAMIQUE

La notion de réputation vise à pondérer l'influence des clients en fonction de leur fiabilité passée, renforçant ainsi la robustesse du processus d'agrégation, comme l'illustre la Figure 2.10.

— Chaque client se voit attribuer un score de réputation calculé en fonction de la cohérence de ses mises à jour et de leur alignement avec le modèle global.

— Ce score est mis à jour dynamiquement après chaque itération d'entraînement.

— Les clients affichant une faible réputation voient leur contribution réduite, ce qui améliore la résilience du système face aux comportements malveillants.

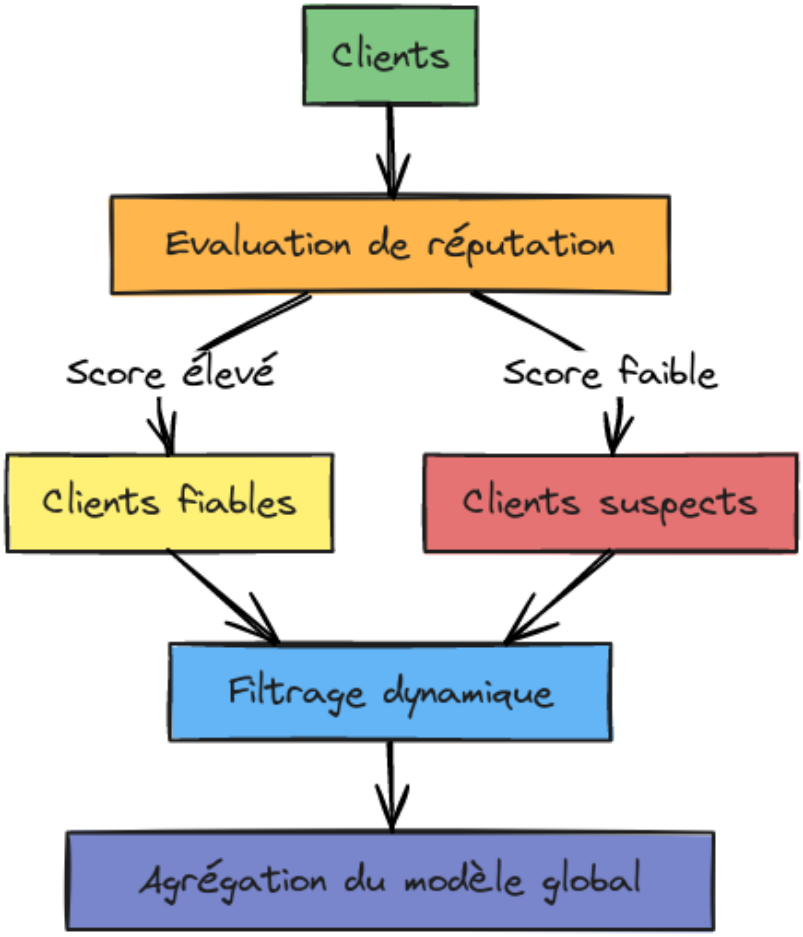

FIGURE 2.10 : Illustration du mécanisme de réputation et de filtrage dynamique dans un système d'apprentissage fédéré. Les clients fiables bénéficient d'une pondération accrue, tandis que les clients suspects voient leur influence réduite.

L'ensemble de ces outils : similarité cosinus, valeur de Shapley, approximations TMC et GTG, et réputation dynamique, constitue le socle théorique de l'approche RBFL. Ils permettent :

— D'identifier les contributions alignées et divergentes.

— D'évaluer la contribution marginale des clients de manière équitable.

— De filtrer ou réduire l'influence des participants malveillants.

Ces notions mathématiques servent de fondation pour l'analyse des travaux existants et pour la conception de mécanismes robustes et équitables dans les systèmes FL, sujet du chapitre suivant.

2.7 COMPARAISON DES PRINCIPAUX ALGORITHMES D'APPRENTISSAGE FÉDÉRÉ

L'agrégation des mises à jour locales est une étape centrale du FL. Selon les scénarios (données homogènes ou fortement hétérogènes, présence de clients malveillants, contraintes de communication), différents algorithmes ont été proposés. Chaque méthode cherche un compromis entre *simplicité*, *robustesse*, *scalabilité* et *équité*.

Chaque algorithme répond à des besoins spécifiques :

- **FedAvg** reste la référence en raison de sa simplicité, mais il est peu robuste aux environnements adversariaux.
- **FedProx** améliore la stabilité en cas de clients hétérogènes.
- **Krum** et **Trimmed Mean** ciblent spécifiquement la résistance aux attaques malveillantes.
- **FedSGD** réduit la communication mais sacrifie la vitesse de convergence.

Ainsi, le choix d'un algorithme dépend étroitement des priorités du système fédéré : *efficacité computationnelle*, *robustesse face aux attaques*, ou encore *équité entre clients*.

2.7.1 DISCUSSION SUR LA ROBUSTESSE ET L'ÉQUITÉ

La comparaison des algorithmes d'agrégation révèle deux grandes tendances. D'un côté, les méthodes simples comme **FedAvg** ou **FedSGD** sont très efficaces lorsque les clients disposent de données homogènes et fiables. Leur faible coût computationnel et communicationnel

en fait des solutions privilégiées dans des contextes peu contraints. Toutefois, leur vulnérabilité aux comportements malveillants et aux distributions non-IID limite leur applicabilité dans des environnements plus réalistes.

À l'opposé, les méthodes dites **robustes** telles que **Krum** ou **Trimmed Mean** introduisent des mécanismes de défense face aux mises à jour aberrantes. Elles améliorent la tolérance aux attaques mais, en contrepartie, peuvent réduire la précision globale lorsque leurs critères de filtrage deviennent trop stricts. Cette rigidité soulève alors la question de l'équilibre entre robustesse et performance.

Enfin, une troisième catégorie d'approches émerge, visant à concilier robustesse et **équité** entre participants. Celles-ci reposent sur l'intégration de mécanismes complémentaires, tels que la réputation dynamique des clients ou l'évaluation de leur contribution via des outils issus de la théorie des jeux (valeur de Shapley et approximations TMC-Shapley). Ces stratégies offrent une alternative prometteuse en permettant d'identifier les mises à jour fiables tout en valorisant les efforts réels de chaque client.

2.7.2 ILLUSTRATION GRAPHIQUE DE L'AGRÉGATION

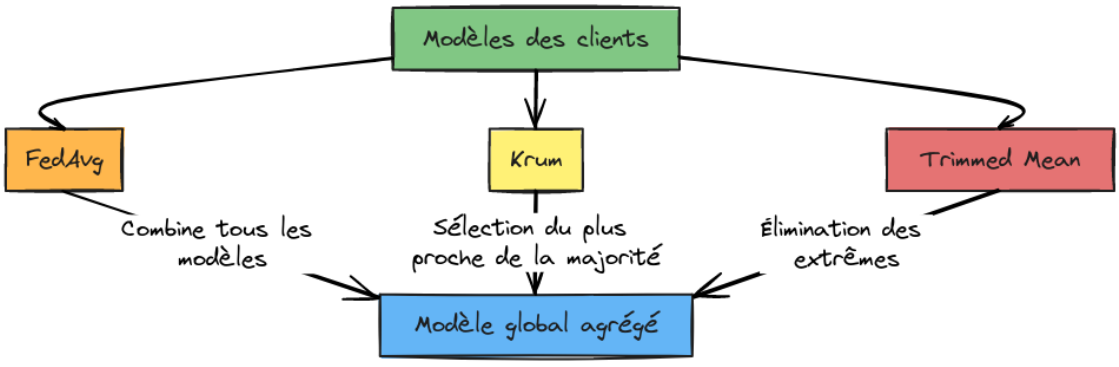

FIGURE 2.11 : Comparaison des logiques d'agrégation

La figure 2.11 illustre de manière schématique les différences conceptuelles entre trois approches représentatives. FedAvg combine sans distinction toutes les mises à jour, Krum sélectionne la mise à jour la plus proche de la majorité, tandis que Trimmed Mean élimine les valeurs extrêmes avant agrégation.

2.7.3 TABLEAU SYNTHÉTIQUE DES DÉFIS ET SOLUTIONS

Les principaux défis rencontrés dans l'apprentissage fédéré et les solutions proposées dans la littérature sont synthétisés dans le tableau 2.9. Ce panorama met en évidence l'articulation entre les limites structurelles du FL (données non-IID, hétérogénéité des clients) et les menaces adversariales (attaques, coûts de communication, risques de confidentialité).

TABLEAU 2.9 : Principaux défis de l'apprentissage fédéré et solutions proposées

Défi	Impact	Solutions proposées
Données non-IID	Divergence du modèle global, biais dans les prédictions	FedProx, personnalisation, regroupement de clients similaires
Hétérogénéité des clients	Convergence lente, instabilité due aux ressources limitées	Ajustement dynamique des rounds, pondération par réputation
Clients malveillants	Biais du modèle global, introduction de backdoors	Krum, Trimmed Mean, évaluation Shapley
Communication coûteuse	Saturation du réseau, latence accrue	FedSGD, compression des gradients, communication périodique
Confidentialité	Risques de fuite de données sensibles	Chiffrement homomorphe, Differential Privacy, techniques FL avancées

2.7.4 CAS CONCRET D'UTILISATION DES ALGORITHMES

Cette analyse comparative met en évidence deux enjeux majeurs : (i) la nécessité de mécanismes **robustes** capables de faire face à des comportements malveillants sans sacrifier

la performance globale, et (ii) l'importance d'une **évaluation équitable** des contributions individuelles afin de garantir une participation durable et loyale.

Ces constats ouvrent la voie à des approches hybrides comme **RBFL**, qui combinent similarité cosinus, valeur de Shapley et gestion dynamique de réputation pour renforcer la sécurité tout en assurant l'équité des contributions.

TRANSITION

L'analyse présentée dans ce chapitre a permis de mettre en lumière les fondements de l'apprentissage fédéré, ses variantes, ainsi que les principaux défis techniques et sécuritaires auxquels il est confronté. Nous avons distingué les modèles locaux et globaux, identifié les limites structurelles de l'apprentissage centralisé, et décrit les apports spécifiques du paradigme fédéré. Une attention particulière a été portée aux algorithmes d'agrégation et à leur capacité à gérer l'hétérogénéité des données, la présence de clients malveillants ou encore les contraintes de communication.

Cette exploration met en évidence deux constats majeurs. D'une part, aucune méthode d'agrégation ne peut, à elle seule, répondre simultanément aux enjeux de robustesse, d'équité et de confidentialité. D'autre part, les mécanismes complémentaires tels que la réputation, l'évaluation contributive (valeur de Shapley et ses approximations) ou encore les techniques de filtrage apparaissent comme des leviers essentiels pour renforcer la résilience des systèmes fédérés.

Ces observations soulignent l'importance de s'appuyer sur un **état de l'art** approfondi afin de comprendre comment la communauté scientifique a tenté de relever ces défis. C'est l'objet du **Chapitre III**, qui présente les principaux travaux connexes portant sur les vulnéra-

bilités, les attaques adversariales et les différentes approches de détection, de défense et de valorisation des contributions en apprentissage fédéré.

CHAPITRE III
TRAVAUX CONNEXES

3.1 INTRODUCTION

L'apprentissage fédéré (*Federated Learning*, FL) a suscité un intérêt considérable au cours de la dernière décennie, notamment en raison de sa capacité à préserver la confidentialité des données tout en permettant la construction de modèles globaux performants. Cependant, ce paradigme collaboratif introduit de nouvelles vulnérabilités liées à la diversité statistique des clients, à l'absence de contrôle direct sur leurs données, ainsi qu'à la possibilité de comportements malveillants ou opportunistes.

Dans ce chapitre, nous présentons un état de l'art structuré autour des axes suivants : (i) les vulnérabilités et menaces inhérentes à l'apprentissage fédéré, (ii) les principales techniques de détection des comportements adverses, (iii) les méthodes d'évaluation et de valorisation des contributions des clients, et enfin (iv) une synthèse critique permettant d'identifier les limites de ces approches et la nécessité d'une solution intégrée telle que **RBFL**.

3.2 VULNÉRABILITÉS ET MENACES DANS L'APPRENTISSAGE FÉDÉRÉ

Les systèmes FL, bien qu'innovants, ne sont pas exempts de vulnérabilités. Ces dernières menacent directement la robustesse, l'équité et la confidentialité des modèles globaux. Elles se regroupent en trois grandes familles.

3.2.1 HÉTÉROGÉNÉITÉ ET DONNÉES NON-IID

Chaque client dispose de données locales différentes en termes de distribution statistique, de volume et de qualité. Cette hétérogénéité (*non-Independent and Identically Distributed*, non-IID) entraîne :

— des difficultés de convergence lors de l'agrégation,

— des biais dans le modèle global favorisant les classes dominantes,

— une complexité accrue pour garantir une évaluation équitable de chaque participant Kairouz *et al.* (2021); Yoo *et al.* (2022).

3.2.2 ATTAQUES VISANT L'INTÉGRITÉ DU MODÈLE

Plusieurs attaques exploitent la confiance accordée aux clients pour corrompre le modèle global :

— **Label-flipping** : inversion ou corruption des étiquettes locales pour biaiser les prédictions globales Khuu *et al.* (2024); Li *et al.* (2022),

— **Backdoor** : introduction de motifs cachés (*triggers*) activant un comportement malveillant du modèle Zhao *et al.* (2024),

— **Gradient poisoning** : perturbation directe des mises à jour envoyées au serveur (par inversion de signe, ajout de bruit ciblé, etc.),

— **Sybil attacks** : création artificielle de multiples clients malveillants pour dominer l'agrégation.

3.2.3 MENACES CONTRE LA CONFIDENTIALITÉ ET LA DISPONIBILITÉ

Au-delà de l'intégrité, les systèmes FL sont aussi exposés à des menaces sur la confidentialité et la disponibilité :

— **Gradient inversion / model inversion** : reconstruction approximative des données locales à partir des gradients,

— **Membership inference** : identification de la présence ou non d'un échantillon donné dans les données d'un client,

— **Attaques par inférence** : extraction d'informations sensibles à partir des mises à jour Lyu *et al.* (2022),

— **Attaques DoS** (Denial of Service) : surcharge du serveur central perturbant l'apprentissage collaboratif.

Ces vulnérabilités démontrent que le FL, malgré ses promesses, nécessite des mécanismes robustes de défense et d'évaluation.

3.3 TECHNIQUES DE DÉTECTION DES COMPORTEMENTS MALVEILLANTS

Pour contrer ces menaces, de nombreuses méthodes ont été proposées. Elles se répartissent en trois grandes catégories complémentaires.

3.3.1 APPROCHES STATISTIQUES ET BASÉES SUR LES GRADIENTS

Ces méthodes examinent les mises à jour locales afin de repérer les comportements anormaux :

— Analyse statistique (ex. PCA) pour identifier des gradients aberrants Khuu *et al.* (2024),

— **Similarité cosinus** pour mesurer la cohérence directionnelle entre gradients locaux et global Lahitani *et al.* (2016),

— Algorithmes d'agrégation robustes : moyenne tronquée, médiane coordonnée, Krum, visant à réduire l'impact des clients malveillants Khraisat *et al.* (2025); Lyu *et al.* (2022).

3.3.2 APPROCHES BASÉES SUR L'ÉVALUATION DE CONTRIBUTION

Ces approches s'appuient sur la théorie des jeux pour mesurer l'apport de chaque client :

— La **valeur de Shapley** assure une répartition équitable de la contribution marginale Ghorbani & Zou (2020); Shapley (1953),

— **TMC-Shapley** réduit le coût computationnel via l'échantillonnage aléatoire Chen *et al.* (2024),

— **GTG-Shapley** intègre une troncature pour accroître la robustesse face aux comportements adverses Liu *et al.* (2021).

3.3.3 APPROCHES BASÉES SUR LA RÉPUTATION

Ces techniques évaluent la fiabilité d'un client à travers son historique :

— Attribution de scores de réputation selon la cohérence et l'honnêteté des mises à jour Kang *et al.* (2019); Xu & Lyu (2020),

— Pondération dynamique de l'influence des clients pour favoriser les contributeurs fiables Lavaur *et al.* (2024),

— Combinaison avec d'autres techniques (par ex. réputation + Shapley, ou réputation + filtrage statistique) pour accroître la résilience.

3.3.4 MÉTHODES HYBRIDES RÉCENTES

Plusieurs travaux proposent de combiner différentes catégories :

— SHAP couplé à un classifieur SVM pour distinguer comportements honnêtes et malveillants,

— Réputation intégrée à des techniques de clustering pour isoler des sous-groupes fiables,

— Défenses basées sur la divergence de distributions (ex. pullback-Leibler divergence) pour caractériser statistiquement les anomalies.

3.4 ÉVALUATION ET VALORISATION DES CONTRIBUTIONS DES CLIENTS

Une étape essentielle en FL consiste à quantifier et valoriser la contribution de chaque client. Au-delà de l'équité, cette valorisation permet de renforcer la sécurité en identifiant les comportements nuisibles.

3.4.1 VALEUR DE SHAPLEY

La valeur de Shapley, issue de la théorie des jeux coopératifs, mesure la contribution marginale moyenne d'un client. Elle satisfait des propriétés d'équité, mais reste coûteuse à calculer pour un grand nombre de participants Ghorbani & Zou (2020, 2019).

3.4.2 APPROXIMATIONS ADAPTÉES AU FL

— **TMC-Shapley** : approximation par échantillonnage de permutations Chen *et al.* (2024); Xu *et al.* (2021),

— **GTG-Shapley** : intégration d'une troncature robuste pour contrer les gradients malveillants Liu *et al.* (2021).

3.4.3 MÉCANISMES DE RÉPUTATION ET FILTRAGE

Les mécanismes de réputation attribuent un score de fiabilité et modulent l'influence d'un client lors de l'agrégation. Cette approche favorise les participants loyaux tout en filtrant les free-riders ou les attaquants persistants Kang *et al.* (2019); Xu & Lyu (2020); Lavaur *et al.* (2024).

3.4.4 APPLICATIONS CONCRÈTES

— **Santé** : contribution équitable des hôpitaux à des modèles prédictifs tout en respectant le secret médical Kumar *et al.* (2022); Li *et al.* (2024),

— **IoT et mobile** : pondération de capteurs selon la qualité des données pour améliorer la robustesse Li *et al.* (2022); Yin & Zeng (2024),

— **Finance** : évaluation des contributions des institutions financières pour la prédiction de risques Yoo *et al.* (2022).

3.5 SYNTHÈSE DU CHAPITRE

L'examen des travaux existants met en lumière la diversité des menaces et la variété des mécanismes de défense proposés.

— Les approches **statistiques** sont rapides et peu coûteuses, mais limitées face à des attaques sophistiquées.

— Les méthodes basées sur la **valeur de Shapley** garantissent une équité théorique, mais posent un problème de scalabilité.

— Les systèmes de **réputation** apportent une régulation dynamique, mais nécessitent un calibrage précis des seuils.

Ces constats révèlent qu'aucune solution prise isolément ne permet de répondre aux enjeux de robustesse, d'équité et de scalabilité. Cela justifie la conception d'approches intégrées. C'est précisément l'objectif de l'approche **RBFL**, qui combine trois dimensions complémentaires : (i) une fonction d'utilité directionnelle basée sur la similarité cosinus, (ii) une approximation TMC-Shapley pour une évaluation équitable, (iii) un mécanisme de réputation dynamique et adaptatif.

Le chapitre suivant présentera en détail la formalisation mathématique de cette approche et son algorithme global.

CHAPITRE IV

APPROCHE RBFL POUR UNE ÉVALUATION ROBUSTE ET ÉQUITABLE EN APPRENTISSAGE FÉDÉRÉ

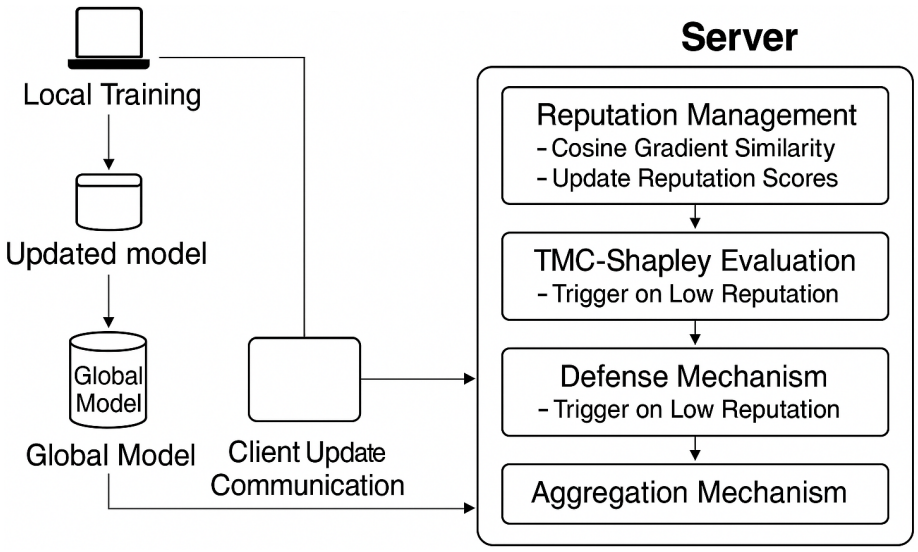

FIGURE 4.1 : Architechture de RBFL : mesure d'utilité (cosine), estimation des contributions (TMC-Shapley), réputation et filtrage réactif, intégrés au cycle fédéré.

4.1 INTRODUCTION

Le succès de l'apprentissage fédéré repose sur la participation active et honnête d'un grand nombre de clients distribués. Toutefois, dans des environnements réels, plusieurs défis émergent : (i) certains clients peuvent se comporter de manière malveillante, par exemple en injectant des mises à jour empoisonnées pour dégrader volontairement le modèle global ; (ii) d'autres peuvent adopter un comportement opportuniste (*freeriders*), ne contribuant que faiblement à l'effort collectif ; (iii) enfin, la qualité des données locales varie fortement d'un client à l'autre, ce qui rend difficile une évaluation équilibrée des contributions.

Les méthodes classiques d'agrégation comme *FedAvg* supposent une participation honnête et homogène des clients. Cette hypothèse, trop restrictive, fragilise les performances dans des environnements hostiles. Il devient donc nécessaire de concevoir des mécanismes capables de distinguer les contributions fiables des comportements nuisibles, en garantissant à la fois robustesse et équité.

L'approche **RBFL** (*Reputation-Based Federated Learning*) renforce la robustesse et l'équité de l'apprentissage fédéré en combinant trois briques complémentaires :

— une **fonction d'utilité directionnelle** fondée sur la similarité cosinus des gradients ;

— une **évaluation équitable des contributions** par approximation **TMC-Shapley** ;

— une **gestion dynamique de la réputation** et un **filtrage réactif** des clients .

La Figure 4.1 présente l'architecture logique de RBFL ainsi que les interactions entre ses composants.

4.2 CADRE FORMEL DE L'APPRENTISSAGE FÉDÉRÉ

Considérons l'ensemble des clients $\mathcal{N} = \{1, \ldots, n\}$. Chaque client i dispose d'un jeu de données local D_i de taille $n_i = |D_i|$. L'objectif est d'optimiser un modèle global w en minimisant la fonction de perte pondérée suivante :

$$F(w) = \sum_{i=1}^{n} p_i F_i(w), \quad \text{où} \quad p_i = \frac{n_i}{\sum_{j=1}^{n} n_j}$$

avec $F_i(w)$ la fonction de perte locale du client i.

Le protocole fédéré suit un schéma itératif composé des étapes suivantes à chaque round t :

1. Le serveur sélectionne un sous-ensemble de clients et diffuse le modèle global w_t.
2. Chaque client entraîne localement w_t sur ses données D_i, produisant une mise à jour Δw_i par descente de gradient :

$$w_{t+1}^i = w_t - \eta \nabla F_i(w_t)$$

3. Les clients transmettent leurs mises à jour locales au serveur.
4. Le serveur agrège ces mises à jour pondérées :

$$w_{t+1} = w_t - \eta \sum_i p_i \Delta w_i$$

Cette procédure, connue sous le nom de *FedAvg* McMahan *et al.* (2016a), est simple et efficace, mais vulnérable aux comportements déviants, justifiant l'introduction de mécanismes robustes tels que RBFL.

4.3 FONCTION D'UTILITÉ BASÉE SUR LA SIMILARITÉ COSINUS DES GRADIENTS

Afin d'évaluer la qualité d'une coalition de clients $S \subseteq \mathcal{N}$, RBFL définit la fonction d'utilité $v(S)$ à partir de la **similarité cosinus** entre le gradient global $\nabla F(w_t)$ et la somme des gradients des clients de S :

$$v(S) = \text{Cosine}\left(\nabla F(w_t), \sum_{i \in S} \nabla F_i(w_t)\right)$$

$$\text{Cosine}(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

Cette mesure, comprise entre -1 et 1, indique si l'effet conjoint d'une coalition va dans la même direction que le gradient global. Un client malveillant enverra typiquement un gradient opposé, donnant une valeur négative facilement détectable.

Ainsi, la similarité cosinus constitue une base robuste et intuitive pour juger de l'alignement des contributions.

4.4 ÉVALUATION ÉQUITABLE VIA APPROXIMATION TMC-SHAPLEY

4.4.1 RAPPEL SUR LA VALEUR DE SHAPLEY

La **valeur de Shapley** Shapley (1953) attribue à chaque client sa contribution moyenne marginale, calculée sur toutes les permutations possibles des coalitions. Elle garantit équité, symétrie et additivité, mais son coût computationnel ($O(n!)$) la rend impraticable dans des systèmes à grande échelle.

4.4.2 APPROXIMATION TMC-SHAPLEY

Pour surmonter cette limite, RBFL utilise l'approximation **Truncated Monte Carlo Shapley** (TMC-Shapley) Li *et al.* (2024), qui procède par échantillonnage aléatoire de permutations avec arrêt anticipé dès que les contributions deviennent négligeables.

Cette approximation conserve les propriétés d'équité de Shapley, tout en restant scalable pour un grand nombre de clients. Elle permet d'évaluer la valeur relative des participants de manière réaliste et efficace.

4.5 GESTION DYNAMIQUE DE LA RÉPUTATION

La réputation joue un rôle central dans RBFL, car elle reflète le comportement des clients sur la durée.

Pour chaque client i, un score de réputation $R_i^{(t)}$ est mis à jour à chaque round selon la cohérence de ses gradients avec ceux des autres :

$$R_i^{(t+1)} = \alpha R_i^{(t)} + (1-\alpha) \cdot \frac{\text{mean}_j \text{Cosine}(\nabla F_i, \nabla F_j)}{\text{mean}_{j,k} \text{Cosine}(\nabla F_j, \nabla F_k)}$$

avec $\alpha \in (0,1)$ contrôlant l'influence de l'historique.

Les clients ayant un score inférieur à un seuil τ sont exclus ou pondérés faiblement lors de l'agrégation. Ce mécanisme permet d'identifier progressivement les comportements nuisibles, tout en préservant la contribution des clients fiables.

4.6 PRÉSENTATION DE L'ALGORITHME RBFL

L'algorithme *Reputation-Based Filtering and Defence* (RBFL), présenté à la Figure 4.2, constitue la pierre angulaire constitue la pierre angulaire de notre approche visant à renforcer la robustesse et l'équité dans l'apprentissage fédéré. Cet algorithme articule une phase d'entraînement fédéré classique avec des mécanismes de filtrage adaptatifs basés sur la réputation dynamique des clients, ainsi qu'une évaluation contributive via une approximation de la valeur de Shapley (TMC-Shapley).

Algorithm 1 Reputation-Based Filtering and Defence

Input: Dataset $\mathcal{D}$; number of clients N; rounds R; epochs E; learning rate η; reputation threshold τ; poison rate ρ; attack start round r_a; number permutations m
Output: Final global model $\mathcal{M}_G$; Shapley values Φ

// Initialization Phase
Initialize $R_i \leftarrow 1.0$ for all $i \in \{1, ..., N\}$; set $\mathcal{E} \leftarrow \emptyset$
Split dataset $\mathcal{D}$ among N clients (iid / non-iid / clustered)
Select $\lfloor \rho \cdot N \rfloor$ malicious clients: $\mathcal{A} \subseteq \{1, ..., N\}$
Initialize global model $\mathcal{M}_G$

// Federated Training Phase
for round $r = 1$ to R **do**
  $\mathcal{P}_r \leftarrow \{i \mid i \notin \mathcal{E}\}$
  for each $i \in \mathcal{P}_r$ **do**
    Set model $\mathcal{M}_i \leftarrow \mathcal{M}_G$
    if $r \geq r_a$ and $i \in \mathcal{A}$ **then**
      Go to Label Poisoning Phase
    else
      Train $\mathcal{M}_i$ on clean D_i for E epochs
    end if
    Compute gradient $G_i \leftarrow \mathcal{M}_i - \mathcal{M}_G$
    Evaluate local accuracy A_i
  end for

// Label Poisoning Attack Phase (only for $i \in \mathcal{A}$)
for each $i \in \mathcal{P}_r \cap \mathcal{A}$ **do**
  Flip labels in a ρ-fraction of D_i randomly
  Re-train $\mathcal{M}_i$ on poisoned D_i for E epochs
  Re-compute $G_i \leftarrow \mathcal{M}_i - \mathcal{M}_G$
  Re-evaluate local accuracy A_i
end for

// Aggregation Phase
Aggregate $\{\mathcal{M}_i\}$ using FedAvg $\Rightarrow \mathcal{M}_G$
Compute cosine similarity matrix S_{ij} between $\{G_i\}$

// Reputation Update Phase
for each $i \in \mathcal{P}_r$ **do**
  $R_i \leftarrow \alpha R_i + (1 - \alpha) \cdot \frac{\text{mean}(S_i)}{\text{mean}(S)}$
end for

// Reactive Filtering Phase
if $r \geq r_a$ and reactive defense enabled **then**
  for each $i \in \mathcal{P}_r$ **do**
    if $i \in \mathcal{A}$ and $|\mathcal{P}_r| - |\mathcal{A}| \geq 4$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{i\}$
    end if
  end for
end if

// Shapley Valuation Phase
Estimate ϕ_i for all i using TMC-Shapley with m permutations

// Logging Phase
Record: global accuracy, reputations $\{R_i\}$, local accuracies $\{A_i\}$, Shapley values $\{\phi_i\}$

// Final Output
end for
return $\mathcal{M}_G$, Φ

FIGURE 4.2 : Algorithme RBFL : ***Reputation-Based Filtering and Defence*****.**

Comme illustré dans la Figure 4.2, l'algorithme procède par itérations successives comprenant :

— La collecte des mises à jour locales des clients actifs au sein du système fédéré.
— Le calcul des similarités directionnelles des gradients afin d'identifier les contributions cohérentes.
— La mise à jour dynamique des scores de réputation des participants selon leur alignement et leur performance.
— Un filtrage réactif excluant progressivement les clients affichant des comportements malveillants ou incohérents.
— La valorisation contributive grâce à une approximation TMC-Shapley, garantissant une évaluation équitable.
— L'agrégation pondérée des mises à jour pour renouveler le modèle global.

Cet enchaînement cyclique permet non seulement d'assurer la robustesse face aux attaques, mais aussi de maintenir une dynamique d'équité et de participation loyale.

La section suivante détaillera la complexité algorithmique et les résultats qualitatifs obtenus lors des premières expérimentations de montée en charge.

4.7 RÉSUMÉ

Ce chapitre a présenté l'approche **RBFL**, qui combine trois briques méthodologiques complémentaires : (i) la similarité cosinus, permettant de juger l'alignement directionnel des gradients ; (ii) l'approximation TMC-Shapley, assurant une estimation équitable et scalable des contributions ; (iii) un mécanisme dynamique de réputation, garantissant une défense adaptative et durable face aux attaques.

Cette synergie permet de dépasser les limites des méthodes classiques et des approches partielles vues dans le chapitre 3, en renforçant à la fois la robustesse et l'équité du processus fédéré.

Le chapitre suivant présentera la mise en œuvre expérimentale et les résultats empiriques qui confirment l'efficacité de RBFL.

CHAPITRE V
MÉTHODOLOGIE D'ÉVALUATION

Ce chapitre présente l'évaluation empirique de l'approche **RBFL** proposée pour renforcer la résilience de l'apprentissage fédéré face aux attaques par empoisonnement de données et à l'hétérogénéité statistique. L'objectif principal est de valider expérimentalement la robustesse, l'équité et l'efficacité globale de notre méthode dans des contextes simulés représentant diverses configurations réalistes.

L'étude s'articule autour de plusieurs volets :

— la description du protocole expérimental et des paramètres de simulation ;

— l'analyse comparative des performances prédictives globales obtenues dans différents scénarios de distribution des données ;

— l'évaluation de la robustesse face à des attaques adversariales (label flipping) ;

— l'étude du comportement dynamique de la réputation ;

— la discussion des résultats, des limites de l'étude et des recommandations pour une mise en œuvre pratique.

Les résultats présentés dans ce chapitre sont issus d'expériences rigoureuses menées sur plusieurs jeux de données de référence (MNIST, Adult, FetalHealth), avec des scénarios réalistes simulant la nature non IID des données, la présence de clients malveillants, et les contraintes du cadre fédéré. Notre approche est également comparée à des méthodes de référence telles que *FedAvg*, *Leave-One-Out*, *Shapley Value* exacte, ainsi qu'une méthode récente basée sur l'intelligence artificielle explicable (SHAP + SVM).

5.1 PROTOCOLE EXPÉRIMENTAL

Cette section décrit le protocole expérimental adopté pour évaluer l'approche RBFL dans divers scénarios de distribution des données, en présence ou non d'attaques adversariales. Elle détaille l'environnement d'exécution, les jeux de données utilisés, les méthodes comparatives sélectionnées, ainsi que les métriques retenues pour l'analyse des résultats.

5.1.1 ENVIRONNEMENT MATÉRIEL ET LOGICIEL

Les expérimentations ont été conduites sur une machine équipée d'un processeur *Intel(R) Core(TM) i7-9700 CPU* et de *32 Go de mémoire vive*. Le serveur central et les clients sont simulés sur la même machine, ce qui est acceptable dans le cadre de cette évaluation car la localisation physique n'influence pas les mesures de performance.

Le modèle de réseau neuronal utilisé est construit avec l'API `Keras Sequential`, et composé des couches suivantes :

— une couche `Flatten` transformant l'entrée en vecteur unidimensionnel ;

— deux couches denses de `128` et `64` neurones, avec activation ReLU ;

— une couche de sortie avec `softmax` adaptée au nombre de classes.

Chaque expérience est répétée cinq fois et les performances sont moyennées pour atténuer les effets du hasard. Le processus s'étend sur 100 cycles de communication fédérée, avec 100 clients disponibles, dont 50% sont sélectionnés aléatoirement à chaque round.

5.1.2 JEUX DE DONNÉES ET PARTITIONS

Trois jeux de données bien établis sont utilisés comme l'illustre le tableau 5.1 :

TABLEAU 5.1 : Jeux de données utilisés

Jeu de données	Taille	Description
MNIST Deng (2012)	70 000 images (60k train, 10k test)	Images de chiffres manuscrits (28×28).
Adult Becker & Kohavi (1996)	48 842 instances	Données tabulaires socio-économiques pour une classification binaire.
FetalHealth de campos *et al.* (2000)	2 126 enregistrements	Données biomédicales de monitoring fœtal réparties en trois classes.

Trois types de distribution de données ont été définis, illustrant différents degrés d'hétérogénéité :

— **S1 – IID** : répartition aléatoire équitable des données d'entraînement et de test entre les clients.

— **S2 – Non-IID (balanced)** : chaque client reçoit un mélange restreint de classes (simulant la non-IID modérée).

— **S3 – Label-clustered** : chaque client ne possède que des exemples d'une seule classe (non-IID extrême).

5.1.3 SCÉNARIO D'ATTAQUE : EMPOISONNEMENT D'ÉTIQUETTES (LABEL-FLIPPING)

Afin d'évaluer la robustesse de l'approche **RBFL** face aux comportements malveillants, nous avons mis en place un scénario d'attaque par empoisonnement d'étiquettes (*label-flipping*), où certains clients injectent volontairement des erreurs dans leurs données d'entraînement.

L'expérimentation repose sur la configuration suivante tableau 5.2.

TABLEAU 5.2 : Paramètres expérimentaux utilisés pour l'évaluation

Paramètre	Valeur
Jeu de données	FETAL
Nombre total de clients	10
Déclenchement de l'attaque	Après 70 rondes
Nombre d'attaquants	6 (sélection aléatoire)
Type d'attaque	Label-flipping local
Taux d'empoisonnement	0,7 % des données locales

L'objectif de ce scénario est de perturber subtilement l'apprentissage du modèle global, en modifiant une faible fraction des étiquettes afin de rester discret tout en compromettant la performance finale.

Deux variantes expérimentales sont analysées :

1. **Avec mécanisme de filtrage RBFL** : les scores de réputation sont utilisés pour détecter et exclure progressivement les clients suspects du processus d'agrégation.
2. **Sans filtrage** : les clients malveillants ne sont jamais détectés et participent librement à chaque ronde.

Ce protocole permet d'observer l'impact de l'attaque sur la convergence du modèle global et de mesurer l'efficacité du filtrage réactif fondé sur la réputation dynamique.

5.1.4 MÉTHODES DE COMPARAISON ET MÉTRIQUES

Les performances de RBFL sont comparées avec plusieurs méthodes de référence :

— **FedAvg** McMahan *et al.* (2016b) : agrégation simple par moyenne pondérée des gradients ;

— **Leave-One-Out (LOO)** Kearns & Ron (1999) : approche naïve pour estimer la contribution d'un client en le retirant ;

— **Shapley Value exacte (SV)** Wang *et al.* (2020) : méthode rigoureuse mais coûteuse pour quantifier l'utilité marginale ;

— **SHAP + SVM** Khuu *et al.* (2024) : méthode récente basée sur l'explicabilité pour détecter les clients malveillants.

Les métriques principales utilisées sont :

— **Accuracy globale** : taux de bonne classification du modèle agrégé ;

— **Résilience à l'attaque** : écart de précision avant et après attaque ;

— **Équité contributive** : mesure de la distribution des poids attribués aux clients ;

— **Temps d'exécution** : évaluation du coût de calcul de chaque méthode.

5.2 QUESTIONS DE RECHERCHE

Afin d'orienter l'analyse expérimentale et de structurer l'interprétation des résultats, nous avons formulé deux questions de recherche principales. Ces questions visent à évaluer l'efficacité, la robustesse et la faisabilité de l'approche RBFL dans différents contextes.

RQ1 – ROBUSTESSE FACE AUX ATTAQUES ADVERSARIALES

Dans quelle mesure l'approche RBFL est-elle capable de préserver la performance du modèle global en présence d'attaques par empoisonnement d'étiquettes (label-flipping) dans un contexte non-IID ?

Cette question vise à évaluer la résilience du mécanisme de filtrage par réputation intégré à RBFL. L'objectif est de mesurer la capacité du système à atténuer les effets néfastes des clients malveillants sur l'apprentissage collaboratif.

RQ2 – SCALABILITÉ ET COÛT COMPUTATIONNEL

Quelle est la faisabilité pratique de l'approche RBFL en termes de complexité algorithmique et de coût d'exécution, comparativement aux autres méthodes ?

Cette question explore la scalabilité de la méthode, notamment l'impact du nombre de clients, la durée de calcul des scores de contribution, et la viabilité de l'implémentation dans des environnements réels ou contraints en ressources.

5.3 RÉSULTATS EXPÉRIMENTAUX

Cette section présente et analyse les résultats obtenus à travers différentes configurations de distribution des données et scénarios d'évaluation. Nous comparons notre approche **RBFL** (Reputation-Based Federated Learning) avec les méthodes de référence *FedAvg*, *Leave-One-Out (LOO)* et *Shapley Value (SV)*, en mettant l'accent sur l'impact de la distribution des données, la robustesse face aux attaques par empoisonnement, ainsi que l'efficacité de l'intégration de TMC-Shapley.

5.3.1 PERFORMANCE PRÉDICTIVE GLOBALE

Le tableau 5.3 présente les précisions moyennes obtenues par le modèle global après 100 rounds d'entraînement dans trois scénarios de distribution : **S1 – IID**, **S2 – Non-IID**, et **S3 – Single Label**. Nous utilisons trois jeux de données standards : MNIST, Adult et FetalHealth.

Dans un environnement **IID** (S1), *FedAvg* obtient les meilleures performances, confirmant son efficacité en l'absence d'hétérogénéité. En revanche, lorsque les données deviennent **non-IID** (S2), la précision de FedAvg chute drastiquement, révélant sa sensibilité à la variabilité des données locales. Dans ce contexte, **RBFL (TMC)** démontre une nette supériorité avec

TABLEAU 5.3 : Performance comparée de RBFL face aux approches de référence (Accuracy)

Méthode	MNIST			ADULT			FETAL_HEALTH		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
FedAvg	**97.17%**	72.94%	70%	**91.21%**	72.39%	73%	**92.30%**	79.73%	78.30%
LOO	97%	90%	88%	86.01%	85.21%	85.24%	89.91%	89.20%	89.90%
SV	95%	89%	90%	80%	77%	71%	92.09%	**91.55%**	**92.25%**
RBFL (SV)	97.11%	70.06%	96.77%	85.43%	85.38%	85.40%	88.50%	84.52%	89.20%
RBFL (TMC)	95.30%	**91.97%**	**96.81%**	85.72%	**85.88%**	**85.84%**	87.80%	88.17%	89.26%

91.97% pour MNIST et 85.88% pour Adult, grâce à la pondération par réputation et à la prise en compte directionnelle des contributions.

Le scénario **S3 – Single Label**, simulant un cas extrême de non-IID, accentue les écarts entre les méthodes. RBFL (TMC) maintient une excellente robustesse avec des précisions proches de celles du scénario IID, validant l'utilité du filtrage par réputation. La méthode SV, bien que plus coûteuse, conserve un avantage sur de petits jeux de données (ex. FetalHealth) où notre approche nécessite davantage d'itérations pour converger.

5.3.2 ROBUSTESSE FACE AUX ATTAQUES D'EMPOISONNEMENT

Pour tester la résilience de RBFL, nous avons injecté une attaque par *label-flipping* après 70 rounds sur le dataset FetalHealth en scénarios S2 et S3. Six clients malveillants ont modifié une portion de leurs étiquettes afin de dégrader l'apprentissage global.

Les Figures 5.1 et 5.2 comparent la performance de RBFL avec et sans filtrage de réputation. Sans défense, la précision chute brutalement après l'attaque. Avec filtrage, le modèle parvient à restaurer une grande partie de la précision initiale.

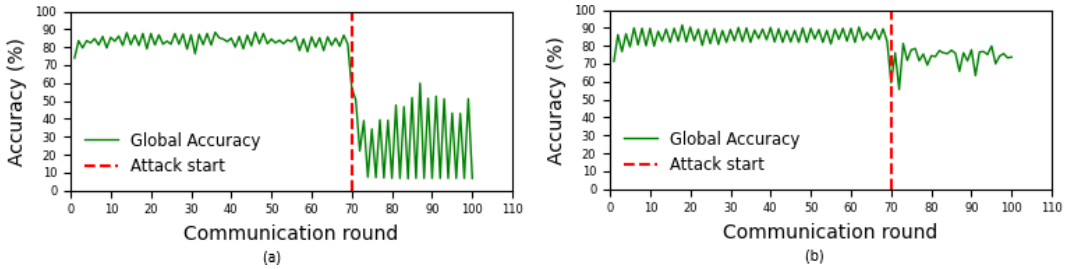

FIGURE 5.1 : Robustesse de RBFL face à l'attaque (Non-IID S2, FetalHealth).

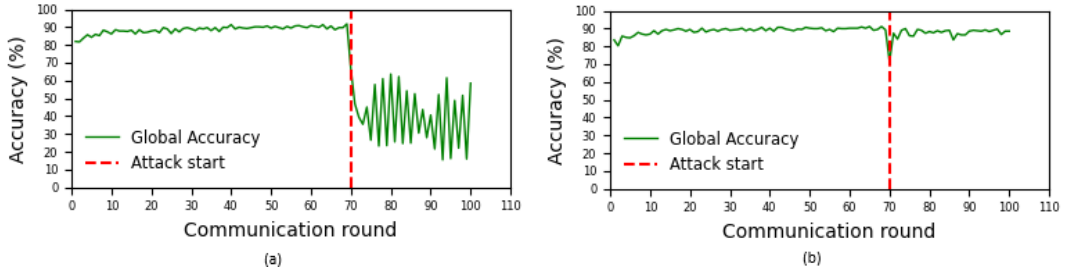

FIGURE 5.2 : Robustesse de RBFL face à l'attaque (Non-IID S3, FetalHealth).

Dans S2, RBFL atteint 88% de précision post-attaque, contre 70% pour la version sans défense. Dans le scénario extrême S3, la précision est quasiment restaurée à 90%, illustrant l'efficacité du mécanisme dynamique de réputation, capable de détecter les comportements incohérents sur plusieurs rounds.

5.3.3 COMPARAISON AVEC UNE MÉTHODE SHAP BASÉE SUR L'EXPLICABILITÉ

Nous avons comparé RBFL à une méthode récente de détection d'attaques par SHAP+SVM proposée dans Khuu *et al.* (2024). Celle-ci repose sur l'utilisation de SHAP values pour détecter des anomalies dans les mises à jour des clients. Les figures 5.3 et 5.4 montrent ses performances avec et sans élimination des clients malveillants.

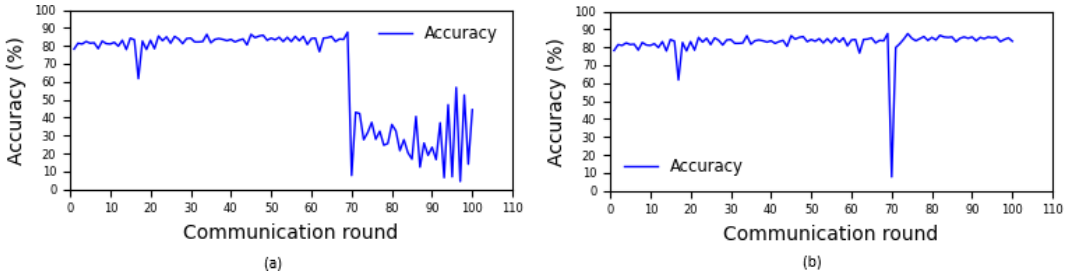

FIGURE 5.3 : Robustesse SHAP face à l'attaque (Non-IID S2, FetalHealth).

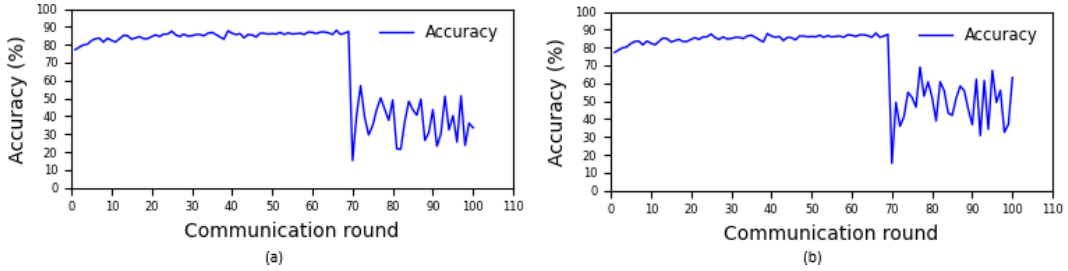

FIGURE 5.4 : Robustesse SHAP face à l'attaque (Non-IID S3, FetalHealth).

Dans S2, la méthode SHAP performe légèrement mieux que RBFL. En revanche, dans S3, son efficacité chute drastiquement, car elle dépend d'un classifieur SVM entraîné sur des vecteurs explicatifs, difficilement représentatifs dans un contexte de données fortement non-IID. RBFL, bien qu'ayant une latence de détection, offre une meilleure stabilité à long terme, car il n'exige pas de données de référence supplémentaires et repose sur des métriques dynamiques internes.

5.4 ÉVALUATION DE LA SCALABILITÉ ET DU COÛT

Cette section analyse la capacité de l'approche **RBFL** à s'adapter à des environnements fédérés de grande échelle tout en maintenant une charge computationnelle maîtrisée. Nous présentons d'abord une analyse de la *complexité algorithmique* des principaux modules, puis des *observations qualitatives* provenant de simulations comportant entre 100 et 300 clients

actifs. Les limites matérielles actuelles ne permettant pas d'aller au-delà sans perturbations d'exécution, une évaluation à plus grande échelle est réservée aux travaux futurs.

5.4.1 COMPLEXITÉ ALGORITHMIQUE

L'approche RBFL repose sur trois composants principaux : (i) l'agrégation FedAvg, (ii) l'évaluation directionnelle des mises à jour via la similarité cosinus, (iii) l'estimation de contribution par la méthode TMC-Shapley.

En notant k le nombre de clients actifs par ronde ($k \approx C \cdot N$), et d la dimension du modèle global, les coûts associés à chaque module se résument comme suit :

— **FedAvg (communication et agrégation)** : Le serveur reçoit puis agrège les mises à jour locales. La complexité reste linéaire :

$$\mathcal{O}(kd).$$

— **Similarité cosinus directionnelle** : Chaque mise à jour locale g_i est comparée à un gradient de référence robuste (p. ex. moyenne tronquée). Contrairement à une comparaison pair-à-pair ($\mathcal{O}(k^2 d)$), la version directionnelle conserve :

$$\mathcal{O}(kd).$$

— **Mise à jour de la réputation (EMA)** : La réputation étant un scalaire par client, le coût est négligeable :

$$\mathcal{O}(k).$$

— **Shapley value exacte** : La complexité théorique :

$$\mathcal{O}(n!)$$

la rend inutilisable dans des environnements fédérés.

— **TMC-Shapley (version utilisée dans RBFL)** : La méthode repose sur m permutations Monte Carlo. En notant $\bar{L} \approx k/2$ la longueur moyenne des coalitions, et c_{util} le coût de la fonction d'utilité (cosinus ou mini-validation), on obtient :

$$\mathcal{O}(m \cdot \bar{L} \cdot c_{\text{util}}).$$

Le mécanisme d'*arrêt anticipé* réduit considérablement le coût en éliminant les permutations non informatives lorsque les contributions marginales se stabilisent.

Synthèse de la complexité. Dans RBFL, le coût par ronde est dominé par la composante linéaire $\mathcal{O}(kd)$ issue de FedAvg et de la similarité cosinus. Le module TMC-Shapley ajoute un surcoût contrôlable via le choix du nombre de permutations et l'arrêt anticipé. Ainsi, l'approche reste scalable pour des centaines de clients, en permettant un compromis ajustable entre précision de l'estimation de contribution et coût computationnel.

5.4.2 OBSERVATIONS QUALITATIVES (100 À 300 CLIENTS)

Nous avons simulé l'approche RBFL en faisant varier le nombre de clients actifs entre 100 et 300. Les observations principales sont les suivantes :

— **Croissance quasi linéaire du temps par ronde.** Le temps d'exécution par ronde augmente proportionnellement à k, conformément aux modèles de complexité. Cette croissance est attendue et liée à l'accumulation des mises à jour locales.

— **Efficacité de la similarité cosinus.** L'utilisation d'un gradient de référence robuste permet d'éviter les comparaisons quadratiques entre clients. Ainsi, même à 300 clients, aucune explosion combinatoire n'est observée.

— **Coût maîtrisé du TMC-Shapley.** Les mécanismes d'arrêt anticipé rendent l'estimation des contributions réalisable en pratique. Le nombre effectif de permutations utilisées diminue lorsque les contributions deviennent rapidement stables.

— **Robustesse du modèle global.** L'augmentation du nombre de clients n'entraîne pas de dégradation significative des performances finales. RBFL maintient une précision stable, même dans un environnement très hétérogène.

Synthèse. Ces résultats montrent que l'approche RBFL se comporte de manière cohérente avec son analyse théorique : elle demeure scalable, flexible et robuste, même lorsque le nombre de clients actifs augmente fortement.

5.4.3 PERSPECTIVES

Pour consolider ces résultats, des expérimentations sur des infrastructures plus puissantes sont prévues. Elles permettront d'étudier la scalabilité au-delà du millier de clients, en évaluant notamment :

— l'évolution du temps total d'apprentissage,

— la consommation mémoire associée aux gradients,

— l'efficacité de techniques supplémentaires d'optimisation (sous-échantillonnage, hiérarchisation des filtres cosinus, compression des mises à jour).

5.5 DISCUSSION INTÉGRÉE

Cette section discute des principaux enseignements issus de notre étude expérimentale et met en perspective les apports de notre approche RBFL comparée aux méthodes existantes.

5.5.1 APPROXIMATION DE LA VALEUR DE SHAPLEY PAR TMC-SHAPLEY

La valeur de Shapley exacte offre une évaluation équitable des contributions, mais sa complexité computationnelle (en $O(n!)$) la rend inapplicable à grande échelle. Pour pallier cela, nous utilisons l'approximation TMC-Shapley, qui repose sur des permutations Monte Carlo et réduit considérablement le coût sans compromettre significativement la qualité de l'évaluation.

Les résultats obtenus sur le jeu de données *Adult* montrent que TMC-Shapley atteint une précision comparable à celle de SV tout en nécessitant 45% de temps de calcul en moins par round (cf. Fig. 5.5 et 5.6). De plus, la convergence est plus rapide (dès le round 4 pour TMC contre le round 7 pour SV). Cela valide l'efficacité et la pertinence de l'approximation dans un contexte d'apprentissage fédéré à grande échelle.

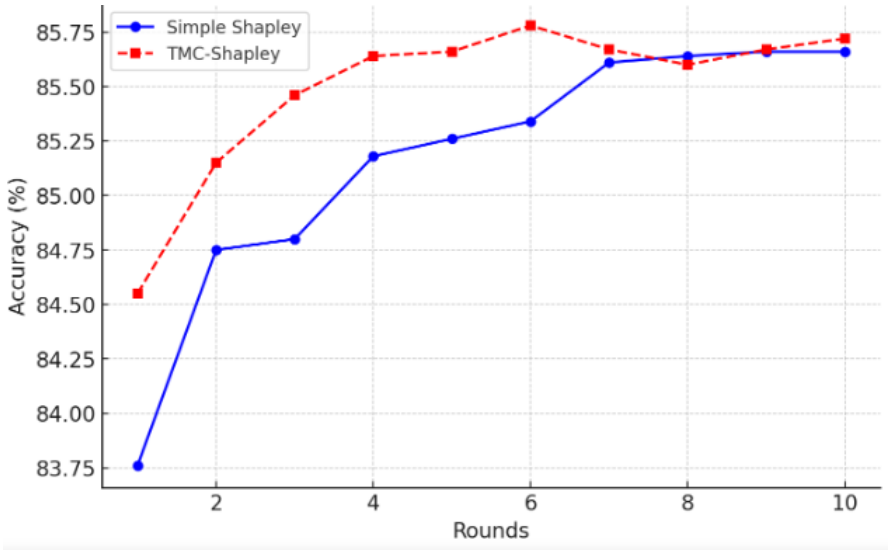

FIGURE 5.5 : Comparaison de la précision entre Shapley simple et TMC-Shapley (Adult).

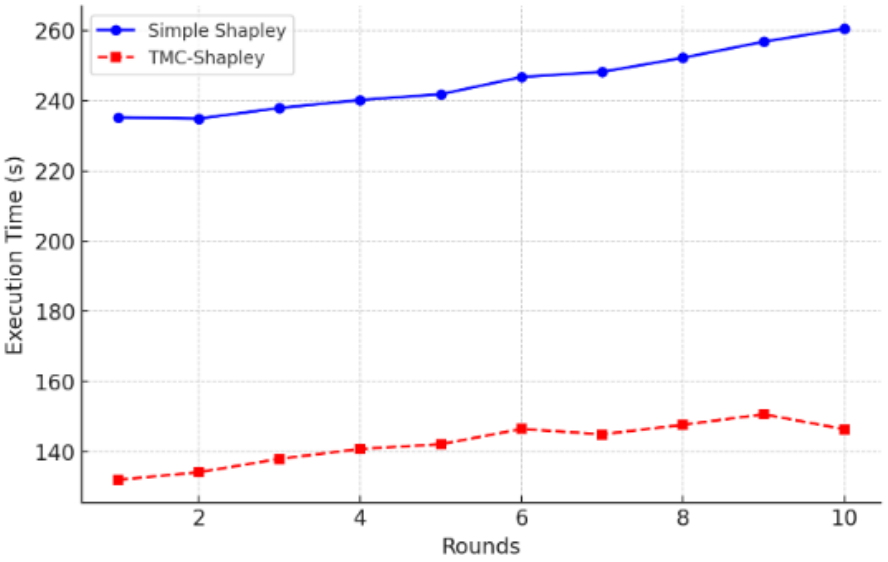

FIGURE 5.6 : Comparaison du temps de calcul entre Shapley simple et TMC-Shapley (Adult).

5.5.2 IMPACT DU FILTRAGE PAR RÉPUTATION

Les résultats montrent que le mécanisme de réputation intégré dans **RBFL** joue un rôle déterminant dans la restauration et la stabilisation des performances du modèle global. Lors d'attaques par *label-flipping*, la réputation permet d'identifier progressivement les clients malveillants et de réduire leur influence. Cette réduction d'impact se traduit par une remontée systématique de la précision après la chute initiale induite par l'attaque (voir Fig. 5.1 et Fig. 5.2).

Un point sensible concerne le choix du seuil de réputation τ. Une recherche sur grille dans l'intervalle $[0.70, 0.95]$ a permis d'identifier des valeurs optimales propres à chaque jeu de données : $\tau = 0.85$ pour *MNIST* et *Adult*, et $\tau = 0.80$ pour *FetalHealth*. Ces seuils assurent un filtrage efficace des mises à jour adverses, tout en préservant la diversité statistique essentielle au maintien d'un modèle global fiable, notamment en configuration non-IID.

5.5.3 DÉPENDANCE AUX CARACTÉRISTIQUES DES JEUX DE DONNÉES

Les performances de RBFL varient selon la nature et la complexité statistique des jeux de données utilisés. Les observations principales sont les suivantes :

— **MNIST** : la distribution est relativement homogène entre clients, ce qui permet aux méthodes classiques (FedAvg, SV) d'obtenir de bons résultats. RBFL conserve son avantage mais l'écart est modéré.

— **Adult** et **FetalHealth** : ces jeux présentent une forte hétérogénéité inter-clients. Dans ces contextes, l'approche RBFL se distingue par une meilleure capacité à stabiliser l'apprentissage et à atténuer les effets négatifs des mises à jour bruitées ou malveillantes.

Ainsi, l'approche montre une adaptabilité supérieure aux environnements complexes où les méthodes standards peinent à maintenir la cohérence du modèle global.

5.5.4 SCALABILITÉ DE L'APPROCHE

L'analyse met en évidence que RBFL demeure scalable grâce à trois facteurs principaux :

— **Coût linéaire du filtrage directionnel et de la réputation.** Le calcul de la similarité cosinus et la mise à jour de la réputation s'effectuent en $\mathcal{O}(n)$, ce qui permet de traiter efficacement un grand nombre de clients.

— **Parallélisation naturelle.** Les calculs liés aux gradients locaux, aux scores de réputation et aux utilités TMC-Shapley sont indépendants d'un client à l'autre, facilitant leur exécution en parallèle.

— **TMC-Shapley paramétrable.** Le recours à une approximation Monte Carlo permet d'adapter dynamiquement le coût computationnel en ajustant le nombre de permutations utilisées. L'arrêt anticipé réduit encore ce coût en détectant la stabilisation précoce des contributions marginales.

Pour les très grandes fédérations (plusieurs milliers de clients), une extension naturelle consiste à introduire une agrégation hiérarchique par groupes ou clusters de clients similaires, ce qui réduirait encore la charge d'évaluation individuelle.

5.5.5 ADAPTATION DYNAMIQUE DU TAUX D'APPRENTISSAGE

RBFL inclut un mécanisme d'adaptation dynamique du taux d'apprentissage basé sur le nombre de clients réellement actifs à chaque ronde. Ce mécanisme contribue à stabiliser le processus fédéré en :

— augmentant légèrement le taux d'apprentissage lorsque peu de clients participent, afin de compenser la faible diversité statistique des mises à jour ;

— réduisant ce taux lorsqu'un grand nombre de clients sont actifs, ce qui limite les instabilités induites par des contributions hétérogènes ou potentiellement adverses.

Contrairement à FedAvg, qui repose sur un taux fixe tout au long de l'entraînement, RBFL maintient une progression plus régulière et mieux contrôlée, même en présence de fluctuations de participation et de variations importantes dans les mises à jour locales.

5.6 LIMITES ET VALIDITÉ DE L'ÉTUDE

Cette section analyse de manière approfondie les limites méthodologiques et les menaces à la validité de l'étude. Elle vise à examiner dans quelle mesure les choix conceptuels, expérimentaux et pratiques peuvent influencer la robustesse des conclusions obtenues. L'objectif n'est pas d'affaiblir les résultats, mais de situer clairement la portée scientifique du travail et de mettre en évidence les précautions nécessaires lors de l'interprétation ou de la généralisation des résultats.

5.6.1 VALIDITÉ CONCEPTUELLE

La validité conceptuelle fait référence à la précision avec laquelle les indicateurs utilisés reflètent les concepts théoriques étudiés. Dans notre cas, nous avons utilisé l'approximation TMC-Shapley pour estimer la contribution des clients, et la similarité cosinus pour calculer la réputation. Bien que ces mesures soient justifiées dans la littérature, elles peuvent ne pas saisir entièrement la qualité sémantique des contributions. De plus, le choix d'un seuil fixe pour la réputation peut ne pas refléter toutes les dynamiques de confiance entre clients, notamment dans des contextes plus complexes ou des réseaux très dynamiques.

5.6.2 VALIDITÉ INTERNE

La validité interne concerne la capacité à établir un lien de causalité entre les mécanismes proposés (RBFL) et les résultats obtenus. Dans notre cadre expérimental simulé, nous avons contrôlé plusieurs facteurs (nombre de clients, type de distribution) afin d'isoler l'effet de notre mécanisme de filtrage basé sur la réputation. Toutefois, l'absence de variabilité dans les comportements adverses et l'hypothèse que les clients malveillants agissent de manière constante sur plusieurs rounds peuvent ne pas refléter fidèlement les comportements adaptatifs d'attaquants réels.

5.6.3 VALIDITÉ EXTERNE

La validité externe fait référence à la généralisation des résultats à d'autres environnements ou contextes. Nos expériences ont été menées sur des jeux de données bien connus (MNIST, Adult, FetalHealth) avec une simulation logicielle sur une seule machine. Cette configuration limite la généralisabilité des résultats à des systèmes FL déployés en conditions réelles avec une distribution géographique, des connexions intermittentes, ou des types de données plus complexes (textuelles, audio, multimodales). De plus, l'attaque simulée est de type label flipping (ce qui est relativement simple) par rapport à d'autres attaques relativement plus sofistiqués et aussi par rapport aux stratégies d'évasion avancées.

5.6.4 LIMITATIONS PRATIQUES

Au-delà des considérations théoriques, plusieurs limites d'ordre pratique ont été identifiées lors de la mise en œuvre du cadre RBFL :

— **Coût computationnel :** Même si TMC-Shapley réduit drastiquement le coût de calcul par rapport au Shapley exact, il demeure relativement coûteux lorsque le nombre de clients augmente ou lorsque le modèle global est profond. Dans un système réel à grande échelle, cette étape pourrait devenir un goulot d'étranglement.

— **Réglage du seuil de réputation :** La valeur τ doit être ajustée empiriquement selon le jeu de données, le modèle et la sévérité de l'attaque. Ce caractère non automatique pose problème dans les environnements où l'on souhaite une adaptation continue sans supervision humaine.

— **Nature cumulative de la réputation :** Le mécanisme actuel favorise une détection réactive. Un client peut effectuer plusieurs actions malveillantes avant que son score ne tombe en dessous du seuil, ce qui peut être problématique dans les contextes critiques (applications médicales, bancaires, ou industrielles).

— **Hypothèse de synchronisation :** Le modèle repose sur FedAvg synchrone. Dans les systèmes réels, les clients participent de manière asynchrone et peuvent ne pas contribuer à chaque round, ce qui complexifie grandement la gestion de la réputation et peut produire des évaluations biaisées.

En somme, bien que l'approche RBFL présente des résultats prometteurs dans les environnements contrôlés, les limites identifiées justifient la nécessité d'étendre les expériences à des scénarios plus réalistes et dynamiques. Des études futures devraient intégrer des adversaires adaptatifs, des modèles de communication non idéaux, des systèmes asynchrones, ainsi qu'une mise à l'échelle à plusieurs centaines ou milliers de clients afin de valider la robustesse généralisée du cadre proposé.

5.7 RECOMMANDATIONS PRATIQUES ET PERSPECTIVES

Au regard des résultats obtenus et des observations issues de l'étude expérimentale, plusieurs recommandations et pistes de recherche peuvent être formulées afin de guider des travaux futurs et de faciliter le déploiement de l'approche RBFL dans des environnements d'apprentissage fédéré réels.

5.7.1 HYPERPARAMÈTRES ADAPTATIFS

Une analyse plus approfondie de la sensibilité du modèle à ses hyperparamètres, et en particulier au seuil de réputation τ, permettrait de renforcer la valeur pratique de l'approche RBFL. Dans l'état actuel, ce seuil est déterminé manuellement par recherche par grille, une méthode certes efficace en simulation, mais potentiellement coûteuse, difficile à généraliser et sensible aux dynamiques changeantes observées dans des systèmes fédérés déployés à grande échelle.

Pour atténuer ces limitations, des mécanismes d'ajustement automatique pourraient être envisagés afin de permettre une adaptation continue aux signaux fournis par le système. Deux familles de stratégies, complémentaires mais encore exploratoires dans le cadre de RBFL, méritent d'être considérées :

1. Règles heuristiques basées sur les métriques en ligne. Cette première approche repose sur l'observation en temps réel de métriques globales (telles que la précision, la perte ou la variance des gradients) et locales (évolution des scores de réputation). Une augmentation soudaine de la dispersion des gradients ou une chute de l'accuracy pourrait, par exemple, suggérer de relever temporairement τ afin de renforcer le filtrage, tandis qu'une stabilisation prolongée inciterait à le diminuer pour réintégrer davantage de clients fiables. Ces règles ont

l'avantage d'être simples, interprétables et peu coûteuses, ce qui les rend potentiellement compatibles avec des environnements fédérés à grande échelle.

2. Optimisation par stratégies de type multi-armed bandits. Une seconde piste, plus algorithmique, consiste à traiter différentes valeurs possibles de τ comme des actions dans un cadre de type bandit. Des algorithmes tels que UCB (Upper Confidence Bound), Thompson Sampling ou certaines variantes bayésiennes pourraient alors être utilisés pour sélectionner progressivement la valeur offrant le meilleur compromis entre robustesse et stabilité du modèle, en fonction des performances observées au fil des rounds. Cette approche permettrait d'apprendre de manière continue un hyperparamètre potentiellement optimal, sans recourir à une recherche exhaustive, tout en restant sensible aux variations du comportement des clients ou à l'apparition de nouveaux types d'attaques.

Dans l'ensemble, ces mécanismes d'adaptation automatique constituent des pistes prometteuses pour réduire la dépendance au réglage manuel d'hyperparamètres critiques et pour doter RBFL d'une capacité d'adaptation accrue aux dynamiques évolutives des systèmes fédérés réels. Leur exploration expérimentale future contribuerait à consolider davantage la portée pratique du cadre proposé.

5.7.2 EXTENSION AUX MENACES AVANCÉES

L'approche RBFL proposée dans cette étude a été évaluée principalement dans le contexte d'attaques par empoisonnement d'étiquettes. Bien que les résultats obtenus soient encourageants, il reste nécessaire de considérer la manière dont cette approche pourrait se comporter face à des menaces plus avancées. Les scénarios suivants, couramment étudiés dans

la littérature sur la sécurité de l'apprentissage fédéré, constituent des pistes importantes pour de futures investigations.

— **Attaques Sybil** : un adversaire peut générer ou contrôler plusieurs clients afin d'influencer l'agrégation globale. En théorie, un mécanisme de réputation pourrait réduire l'impact de ces attaques en limitant le poids des mises à jour incohérentes. Cependant, l'efficacité réelle de RBFL dans ce cadre reste à valider empiriquement.

— **Collusion entre clients malveillants** : plusieurs attaquants peuvent coordonner leurs mises à jour pour tenter de contourner les défenses. Les techniques fondées sur la similarité des gradients ou sur l'estimation marginale via TMC-Shapley pourraient offrir une certaine résilience, mais ces hypothèses nécessitent des scénarios expérimentaux dédiés pour être confirmées.

— **Attaques adaptatives** : certains adversaires ajustent progressivement leur comportement dans le temps afin de réduire les signaux de détection. Le suivi temporel des scores de réputation pourrait, en principe, permettre d'identifier ces variations subtiles, mais l'évaluation de ce comportement dépasse le cadre expérimental de ce mémoire.

— **Attaques de type backdoor** : ces attaques ciblées insèrent des comportements malveillants spécifiques dans le modèle tout en maintenant des performances globales normales. Les mécanismes fondamentaux de RBFL tels que la similarité cosinus et l'évaluation basée sur TMC-Shapley pourraient contribuer à réduire l'impact de ces attaques, mais leur efficacité dans un tel contexte reste à démontrer expérimentalement.

Ainsi, bien que l'architecture RBFL présente des propriétés qui laissent entrevoir un potentiel d'extension face à ces menaces avancées, une évaluation empirique approfondie demeure nécessaire. Ces pistes constituent des perspectives naturelles pour des travaux ultérieurs visant à renforcer la résilience du cadre proposé.

5.7.3 SCALABILITÉ À TRÈS GRANDE ÉCHELLE

Une analyse plus approfondie des conditions de scalabilité s'avère essentielle pour apprécier la valeur pratique de RBFL dans des environnements fédérés réels. En effet, dans la perspective d'un déploiement sur des réseaux comportant plusieurs milliers de clients, l'approche RBFL pourrait nécessiter certaines optimisations afin de maintenir un coût computationnel raisonnable tout en préservant un niveau de robustesse adéquat. Plusieurs pistes, cohérentes avec les tendances observées dans les systèmes distribués à grande échelle, méritent une attention particulière :

— **Évaluation groupée des contributions.** Les clients présentant des gradients similaires pourraient être regroupés de manière à estimer la contribution à l'échelle du groupe plutôt qu'individuellement. Une telle stratégie est susceptible de réduire le nombre d'évaluations nécessaires et de limiter l'impact de la redondance entre clients partageant des comportements proches.

— **Hiérarchisation des clients.** Un filtrage préliminaire reposant sur des métriques peu coûteuses (par exemple la norme du gradient, la stabilité locale ou un score de réputation simplifié) pourrait permettre d'écarter rapidement les mises à jour manifestement aberrantes. Les évaluations plus coûteuses, telles que celles fondées sur TMC-Shapley, seraient alors réservées à un sous-ensemble restreint de clients potentiellement influents.

— **Réduction adaptative des échantillons TMC.** Le nombre de permutations Monte Carlo pourrait être ajusté dynamiquement en fonction de la stabilité des estimations. Lorsque les valeurs marginales semblent converger, l'algorithme pourrait interrompre automatiquement l'échantillonnage afin de réduire l'empreinte computationnelle sans compromettre la précision.

Ces pistes n'ont pas été évaluées expérimentalement dans le cadre de ce travail, mais elles constituent des directions prometteuses pour améliorer la montée en charge de RBFL dans des environnements massifs. Leur exploration future contribuerait à renforcer la capacité de RBFL à s'adapter aux contraintes opérationnelles des systèmes fédérés à grande échelle.

5.7.4 DÉPLOIEMENT EN CONDITIONS RÉELLES

Pour valider l'applicabilité de RBFL dans des contextes opérationnels, un déploiement pilote dans un environnement réel (par exemple hospitalier, bancaire ou éducatif) apparaît nécessaire. Un tel déploiement permettrait d'évaluer la robustesse de l'approche face à la variabilité des connexions, à l'hétérogénéité des données produites in situ et aux comportements potentiellement imprévisibles des clients. Il impliquerait également de garantir la conformité avec les exigences réglementaires (telles que le RGPD) et d'assurer un niveau suffisant de transparence et d'auditabilité des décisions prises par le mécanisme de réputation.

5.7.5 INTÉGRATION AVEC D'AUTRES DÉFENSES

Enfin, RBFL pourrait être intégré dans un cadre de défense hybride combinant plusieurs familles de mécanismes complémentaires :

— **Méthodes d'agrégation robustes** (par exemple Trimmed Mean, Krum) pour limiter l'impact de mises à jour extrêmes.

— **Techniques de détection fondées sur l'explicabilité** (telles que SHAP ou LIME) afin de mieux interpréter et justifier les décisions de filtrage.

— **Schémas d'apprentissage fédéré asynchrone ou personnalisé** pour améliorer l'adaptation aux fortes hétérogénéités de données et de participation côté client.

Une telle intégration ouvrirait la voie à une défense plus complète et modulable, tout en préservant de bonnes performances en termes de précision et de coût computationnel.

CONCLUSION

Ce mémoire a porté une attention particulière à la sécurisation de l'apprentissage fédéré dans un contexte caractérisé par l'hétérogénéité des données et une possible présence de clients malveillants. Face aux limites des schémas d'agrégation traditionnels, souvent vulnérables à des comportements adversaires et insensibles à la qualité réelle des mises à jour, nous avons proposé une approche novatrice, baptisée **RBFL (Reputation-Based Federated Learning)**, qui combine robustesse, équité et scalabilité.

Cette méthode s'appuie sur trois composantes complémentaires et synergiques : une fonction d'utilité directionnelle fondée sur la similarité cosinus des gradients pour évaluer la pertinence immédiate des contributions locales ; une approximation efficace de la valeur de Shapley par Monte Carlo tronqué (TMC-Shapley), assurant une évaluation équitable et cohérente de chaque contribution ; enfin, un mécanisme adaptatif de gestion de la réputation, permettant d'ajuster dynamiquement le poids des clients selon leur comportement historique.

Les expériences réalisées sur différents jeux de données (MNIST, Adult, FetalHealth), dans des contextes variés de distribution (IID, non-IID, mono-étiquette), confirment la pertinence de cette approche. RBFL maintient une haute précision même en présence d'une forte hétérogénéité, détecte et réduit efficacement les conséquences des attaques par empoisonnement d'étiquettes, et améliore l'équité dans la valorisation des contributions tout en préservant la scalabilité du système. Par ailleurs, la comparaison avec des méthodes récentes basées sur l'explicabilité (telles que SHAP combiné à des classifieurs comme SVM) met en avant la stabilité et la résistance de RBFL face aux situations critiques. Bien que la détection ne soit pas immédiate, la dynamique de réputation engendre une récupération progressive et durable des performances globales.

Pour l'avenir, nous proposons plusieurs pistes d'amélioration : l'intégration de RBFL dans des environnements distribués réels adaptés à des contraintes variables de communication ; l'évaluation de sa robustesse face à des attaques plus sophistiquées, telles que Sybil, les collusions ou les attaques par inférence ; l'ajustement automatique des hyperparamètres clés, notamment les seuils de réputation, en fonction des comportements observés ; et l'extension du cadre à d'autres types de données, incluant les données séquentielles ou non structurées comme le texte ou l'audio. Ces évolutions devraient renforcer la généralisabilité et la transférabilité industrielle de l'approche.

En résumé, ce travail contribue substantiellement à l'avancement de l'apprentissage fédéré sécurisé, en proposant un cadre unifié mêlant utilité directionnelle, évaluation équitable des contributions et filtrage par réputation. RBFL se distingue par sa capacité à rester performant, équitable et scalable dans des conditions réalistes, posant ainsi une base solide pour des déploiements à large échelle de systèmes intelligents distribués.

Valorisation scientifique. Les résultats présentés ont été reconnus par la communauté académique à travers leur acceptation à la conférence *IEEE AICCSA 2025 (International Conference on Computer Systems and Applications)*, renforçant la visibilité, la portée scientifique et la pertinence de l'approche RBFL proposée.

BIBLIOGRAPHIE

Becker, B. & Kohavi, R. (1996). *Adult*. UCI Machine Learning Repository. DOI : https ://doi.org/10.24432/C5XW20.

Chen, Y., Li, K., Li, G. & Wang, Y. (2024). Contributions Estimation in Federated Learning : A Comprehensive Experimental Evaluation. Dans *Proceedings of the VLDB Endowment*, Vol. 17, pp. 2077–2090. VLDB Endowment. doi: 10.14778/3659437.3659459

de campos, D. A., Bernardes, J., Garrido, A., de sá, J. M. & leite and, L. P. (2000). SisPorto 2.0 : A Program for Automated Analysis of Cardiotocograms. *Journal of Maternal-Fetal Medicine*, *9*(5), 311–318.

Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, *29*(6), 141–142. doi: 10.1109/MSP.2012.2211477

Efthymiadis, F., Karras, A., Karras, C. & Sioutas, S. (2024). Advanced Optimization Techniques for Federated Learning on Non-IID Data. *Future Internet*, *16*(10). doi: 10.3390/fi16100370

Ghorbani, A. & Zou, J. (2019). What is your data worth ? Equitable Valuation of Data" by Amirata Ghorbani and James Zou. Dans *International Conference on Machine Learning*.

Ghorbani, A. & Zou, J. (2020). Data Shapley : Equitable Valuation of Data for Machine Learning. Dans *International Conference on Machine Learning*, pp. 2242–2251.

Joshi, M., Pal, A. & Sankarasubbu, M. (2022, 4). Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. *ACM Trans. Comput. Healthcare*. doi: 10.1145/3533708

Kairouz, P., McMahan, H. B., Avent, B. & Bellet, e. a. (2021). Advances and Open Problems in Federated Learning. *arXiv :1912.04977 [cs, stat]*. arXiv : 1912.04977, Repéré le 2021-09-24, à http://arxiv.org/abs/1912.04977

Kang, J., Xiong, Z., Niyato, D., Xie, S. & Zhang, J. (2019). Incentive Mechanism for

Reliable Federated Learning : A Joint Optimization Approach to Combining Reputation and Contract Theory. *IEEE Internet of Things Journal*, pp. 10700–10714.

Kearns, M. & Ron, D. (1999). Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, *11*(6), 1427–1453.

Khraisat, A., Alazab, A., Alazab, M., Jan, T., Singh, S. & Uddin, M. A. (2025, 1). Securing federated learning : a defense strategy against targeted data poisoning attack. *Discover Internet of Things*, p. 16.

Khuu, D.-P., Sober, M., Kaaser, D., Fischer, M. & Schulte, S. (2024). Data Poisoning Detection in Federated Learning. Dans *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, p. 1549–1558., New York, NY, USA. Association for Computing Machinery.

Kumar, S., Lakshminarayanan, A. & et al. (2022). Towards More Efficient Data Valuation in Healthcare Federated Learning Using Ensembling. p. 119–129., Berlin, Heidelberg. Springer-Verlag. doi: 10.1007/978-3-031-18523-6_12

Lahitani, A. R., Permanasari, A. E. & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure : Study case in online essay assessment. Dans *2016 4th International Conference on Cyber and IT Service Management*, pp. 1–6.

Lavaur, L., Lechevalier, P.-M., Busnel, Y., Ludinard, R., Texier, G. & Pahl, M.-O. (2024). *RADAR : Model Quality Assessment for Reputation-aware Collaborative Federated Learning*. Repéré à https://hal.science/hal-04681789

Li, J., Guo, W., Han, X., Cai, J. & Liu, X. (2022). *Federated Learning based on Defending Against Data Poisoning Attacks in IoT*. Repéré à https://arxiv.org/abs/2209.06397

Li, J., He, J., Hu, L. & Wu, C. (2024, 4). A Federated Learning Framework via Decentralized Data Valuation for Chronic Disease Healthcare. Dans *ACM International Conference Proceeding Series*, pp. 60–65. Association for Computing Machinery. doi: 10.1145/3669828.3669837

Liu, Z., Chen, Y., Yu, H., Liu, Y. & Cui, L. (2021). GTG-Shapley : Efficient and Accurate

Participant Contribution Evaluation in Federated Learning. Repéré à http://arxiv.org/abs/2109.02053

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q. & Yu, P. S. (2022). *Privacy and Robustness in Federated Learning : Attacks and Defenses*. Repéré à https://arxiv.org/abs/2012.06337

McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. Repéré à http://arxiv.org/abs/1602.05629

McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. Repéré à http://arxiv.org/abs/1602.05629

Riedel, P., Schick, L., Schwerin, R., Reichert, M., Schaudt, D. & Hafner, A. (2024). Comparative analysis of open-source federated learning frameworks - a literature-based survey and review. *International Journal of Machine Learning and Cybernetics*, pp. 5257–5278. doi: 10.1007/s13042-024-02234-z

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, *2*.

Sharma, P., Choi, K., Krejcar, O., Blazek, P., Bhatia, V. & Prakash, S. (2023). Securing Optical Networks Using Quantum-Secured Blockchain : An Overview. *Sensors*, *23*(3). doi: 10.3390/s23031228

Trend, A. T. (2023, 02). *Cosine Similarity [Image]*. Consulté le 16 septembre 2025, Repéré à https://aitechtrend.com/wp-content/uploads/2023/02/Cosine-similarity.jpg

Wang, T., Rausch, J., Zhang, C., Jia, R. & Song, D. (2020). A Principled Approach to Data Valuation for Federated Learning. Repéré à http://arxiv.org/abs/2009.06192

Xia, Y., Yang, D., Li, W., Myronenko, A., Xu, D., Obinata, H., Mori, H., An, P., Harmon, S., Turkbey, E., Turkbey, B., Wood, B., Patella, F., Stellato, E., Carrafiello, G., Ierardi, A., Yuille, A. & Roth, H. (2021, 04). *Auto-FedAvg : Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation*. doi: 10.48550/arXiv.2104.10195

Xu, X. & Lyu, L. (2020). A Reputation Mechanism Is All You Need : Collaborative Fairness and Adversarial Robustness in Federated Learning. Repéré à http://arxiv.org/abs/2011.10464

Xu, X., Wu, Z., Foo, C. S. & Low, B. K. H. (2021). Validation Free and Replication Robust Volume-based Data Valuation. Dans M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Éds.). *Advances in Neural Information Processing Systems*, Vol. 34, pp. 10837–10848. Curran Associates, Inc.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2019). *Federated Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers. Repéré à https://books.google.ca/books?id=JdPGDwAAQBAJ

Yin, C. & Zeng, Q. (2024). Defending Against Data Poisoning Attack in Federated Learning With Non-IID Data. *IEEE Transactions on Computational Social Systems*, *11*(2), 2313–2325. doi: 10.1109/TCSS.2023.3296885

Yoo, J. H., Jeong, H., Lee, J. & Chung, T.-M. (2022). Open problems in medical federated learning. *International Journal of Web Information Systems*, *18*(2/3), 77–99.

Zhang, J., Zhu, H., Wang, F., Zhao, J., Xu, Q. & Li, H. (2022). Security and Privacy Threats to Federated Learning : Issues, Methods, and Challenges. *Security and Communication Networks*, pp. 1–24. doi: 10.1155/2022/2886795

Zhao, J., Zhu, H., Wang, F., Zheng, Y., Lu, R. & Li, H. (2024). Efficient and Privacy-Preserving Federated Learning Against Poisoning Adversaries. *IEEE Transactions on Services Computing*, *17*(5), 2320–2333. doi: 10.1109/TSC.2024.3377931

Zhu, S., Zeng, J., Wang, S., Sun, Y., Li, X., Yao, Y. & Peng, Z. (2024). *On ADMM in Heterogeneous Federated Learning : Personalization, Robustness, and Fairness*. Repéré à https://arxiv.org/abs/2407.16397