UQAC
Université du Québec
à Chicoutimi

INTEGRATION D'ALGORITHMES POST-QUANTIQUES DANS UN APPRENTISSAGE FÉDÉRÉ : UNE ÉVALUATION EMPIRIQUE POUR LA DÉTECTION D'OBJETS.

PAR MAMADOU ALIOU MAIMOUNA DIALLO

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 (Federated Learning, FL) facilite l'entraînement collaboratif de modèles tout en conservant les données brutes localement, mais les communications et les gradients échangés restent vulnérables aux attaques d'inversion et aux menaces quantiques futures capables de compromettre l'établissement de clés classique. Cette recherche a pour but d'étudier l'intégration pratique des mécanismes cryptographiques post-quantiques dans un pipeline de FL appliqué à la détection d'objets en vision par ordinateur. Le mécanisme de sécurité hybride proposé combine, Kyber512 (KEM) sélectionné par le NIST) pour faire l'encapsulation des clés résistante aux attaquent quantique, Cristal-Dilitium pour l'authentification et l'intégrité des différentes mises à jour, HKDF-SHA256 pour la dérivation de clés symétriques, et AES-CBC pour le chiffrement efficace de tous les gradients et du modèle global. Il intègre également un masquage pairwise des gradients basés sur une topologie de graphe de voisins, garantissant l'annulation des masques lors de l'agrégation sécurisée FedAvg afin que le serveur n'observe jamais les gradients individuels en clair. Une évaluation empirique est réalisée dans un contexte adversarial réaliste pour mesurer la confidentialité des gradients, le surcoût induit par la sécurité, la stabilité de la convergence et les performances de détection. Les différents résultats expérimentaux montrent que l'approche réduit fortement la corrélation entre les gradients et données d'origine, et maintient une convergence stable et ils préservent des performances très proches du FL non sécurisé, tout en renforçant la sécurité à long terme face au risque (harvest-now, decrypt-later). Ce travail contribue à la conception de systèmes FL pratiques, sécuritaires et résilients aux attaquants post-quantiques, adaptés aux modèles profonds de vision par ordinateur, conformément aux standards actuels de la cryptographie post-quantique.

ABSTRACT

Federated Learning (FL) enables collaborative model training while keeping raw data local, but communications and exchanged gradients remain vulnerable to inversion attacks and future quantum threats that can break classical key establishment schemes. This research investigates the practical integration of post-quantum cryptographic mechanisms into an FL pipeline applied to object detection in computer vision. The hybrid security mechanism I proposed combines Kyber-512 (NIST-selected KEM) for quantum-resistant key encapsulation, CRYSTALS-Dilithium for authentication and integrity of different model updates, HKDF-SHA256 for symmetric key derivation, and AES-CBC for efficient encryption of all gradients and the global model. In addition, pairwise gradient masking over a neighbor graph topology ensures that masks cancel out during secure FedAvg aggregation, so the server only sees the masked sum and never any individual client gradient in plaintext. Empirical evaluation is performed in a realistic adversarial setting to measure gradient privacy, security-induced overhead, convergence stability, and detection performance. Experimental results show that the approach significantly reduces gradient-data correlation, maintains stable convergence, preserves object detection performance close to non-secure FL, and strengthens long-term security against "harvest-now, decrypt-later" quantum risks. This work contributes to designing practical, deployable, quantum-resilient FL systems for deep vision models, aligned with current post-quantum cryptography standards.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

AES Advanced Encryption Standard
AES-CBC Advanced Encryption Standard – Cipher Block Chaining
CNN Convolutional Neural Network
CPU Central Processing Unit
CRC Cyclic Redundancy Check
DLG Deep Leakage from Gradients
DP Differential Privacy
ECC Elliptic Curve Cryptography
FedAvg Federated Averaging
FL Federated Learning
GLRS Gradient Leakage Reduction Score
GPU Graphics Processing Unit
HE Homomorphic Encryption
HKDF HMAC-based Key Derivation Function
HMAC Hash-based Message Authentication Code
IDLG Improved Deep Leakage from Gradients
IoT Internet of Things
KEM Key Encapsulation Mechanism
L2 Euclidean Norm
mAP mean Average Precision
mAP50 mean Average Precision at IoU 0.50
mAP5095 mean Average Precision averaged from IoU 0.50 to 0.95
MPC Multi-Party Computation
MSE Mean Squared Error
NIST National Institute of Standards and Technology
NonIID Non-Independent and Identically Distributed
PQC Post-Quantum Cryptography
PSNR Peak Signal to Noise Ratio
RSA Rivest Shamir Adleman
SNR Signal to Noise Ratio
SSDLite Single Shot Detector Lite
SSIM Structural Similarity Index
TLS Transport Layer Security
YOLO You Only Look Once

DÉDICACE

À ALLAH, le Clément et le miséricordieux, le Tout-Puissant, je rends grâce pour sa bienveillance et sa santé, qui m'as aidé dans la réalisation de ce travail.

Je dédie ce modeste travail :

À mon défunt pére que la paix soit sur lui et à ma chère mère, qui n'a jamais cessé de m'encourager et de se sacrifier pour que je puisse surmonter chaque obstacle, depuis ma naissance jusqu'à aujourd'hui. Que Dieu la garde en bonne santé.

REMERCIEMENTS

Je remercie tout d'abord le Tout-Puissant ALLAH, détenteur de toute âme et de tout pouvoir, de m'avoir facilité cette tâche.

Mes remerciements vont ensuite à mon cher père, que j'ai malheureusement perdu en 2007. À ma chère mère, qui, malgré l'absence de mon défunt père, a pourvu à tous mes besoins, m'offrant l'éducation, les soins, le soutien et s'assurant que je ne manque de rien pour atteindre cet objectif.

À mon épouse bien-aimée, pour son amour et ses encouragements, qui m'attend patiemment dans notre pays natal, la Guinée. À mon frère et mes sœurs pour leur soutien et leurs encouragements constants. Je tiens à exprimer ma gratitude pour l'altruisme et les sacrifices sans limites dont ils ont toujours fait preuve malgré les défis de la vie.

Mes sincères remerciements vont également à mon professeur, M. Fehmi Jaafar, mon directeur de recherche et à Mme Haifa nakouri qui m'ont guidé dès mon arrivée, pour le choix de mes cours, ainsi que dans la réalisation de ce mémoire. J'adresse aussi mes chaleureux remerciements à tous mes formateurs et encadreurs, tant en Guinée qu'au Canada, pour leur compétence, disponibilité, et humanité, qui ont largement contribué à ma formation.

Je remercie également tous les professeurs et le personnel administratif du Département d'Informatique et de Mathématiques (DIM) de l'Université du Québec à Chicoutimi (UQAC) qui m'ont permis de venir au Canada pour acquérir de nouvelles connaissances.

Mes remerciements s'étendent également à l'école primaire Hoolo et le lycée Amilcar Cabral de Mamou en Guinée, où j'ai obtenu mon diplôme de baccalauréat en sciences mathématiques, équivalent ici au Diplôme d'Études collégiales. À l'Université Barack Obama Conakry (UBO), en particulier au département de Génie informatique, où j'ai fait mes premières découvertes des outils de l'informatique et obtenu mon diplôme d'ingénieur, me permettant ainsi d'obtenir une admission pour venir étudier au Canada.

J'adresse mes chaleureux remerciements, marqués d'amitié, d'amour et de fraternité, à tous mes oncles, tantes et camarades, proches ou lointains, pour le soutien qu'ils m'ont apporté.

AVANT-PROPOS

Ce mémoire s'inscrit dans le cadre de mes études de Maîtrise en informatique et vise à répondre à une problématique majeure des systèmes d'apprentissage automatique distribués modernes, la protection de la confidentialité des données et des modèles dans les environnements d'apprentissage fédéré. Avec la montée en puissance des applications fondées sur l'intelligence artificielle et l'apprentissage profond, notamment en vision par ordinateur, la nécessité de concilier performance, sécurité et respect de la vie privée est devenue un enjeu central tant dans les milieux académiques que professionnels. L'apprentissage fédéré propose aujourd'hui une alternative prometteuse au paradigme centralisé, en permettant l'entraînement collaboratif des différents modèles sans faire le partage direct des données brutes. Toutefois, plusieurs travaux démontrent que cette approche reste encore vulnérable à des attaques permettant d'exploiter les gradients échangés et pouvant entraîner des fuites d'informations sensibles. Par ailleurs, l'émergence des ordinateurs quantiques remet en question la robustesse à long terme des schémas cryptographiques classiques qui sont utilisés pour sécuriser ces différents systèmes distribués. Dans ce contexte, ce travail a pour but de concevoir, d'implémenter et d'évaluer le mécanisme d'apprentissage fédéré sécurisé qui intègre des mécanismes de cryptographie post-quantique. Le projet combine plusieurs approches complémentaires, notamment le masquage pairwise des gradients, l'ajout de bruit gaussien qui s'annule à l'agrégation, le chiffrement symétrique, les signatures numériques et l'utilisation de schémas post-quantiques tels que CRYSTALS-Kyber. L'ensemble de ces mécanismes est appliqué à un cas concret de vision par ordinateur qui est basé sur un modèle de détection d'objets de type YOLO, dans un environnement fédéré qui est réaliste en utilisant des données potentiellement non-IID. La réalisation de ce mémoire de recherche m'a permis d'approfondir mes connaissances en apprentissage fédéré, en sécurité des systèmes distribués, en cryptographie post-quantique et en vision par ordinateur. Dans une démarche de valorisation scientifique, l'ensemble des travaux et résultats présentés dans cette recherche a conduit à la rédaction et à la soumission d'un article scientifique complet auprès de la conférence internationale SecureComm (EAI), spécialisée dans les communications sécurisées et les systèmes distribués. Cette soumission s'inscrit dans la continuité des objectifs de ce projet de recherche et vise à confronter les contributions proposées à l'évaluation de la communauté scientifique internationale.

INTRODUCTION

L'apprentissage fédéré (FL) est un modèle d'apprentissage automatique décentralisé, qui permet à plusieurs entités de travailler ensemble pour former un modèle global sans partager leurs données brutes. Ainsi, au lieu de centraliser les données sensibles, chaque client effectue un entraînement local et transmet uniquement des mises à jour du modèle, telles que les gradients ou les poids du modèle, à un serveur central chargé de l'agrégation. Cette approche permet de réduire de manière significative les différents risques liés à la confidentialité qui ont déjà été mis en œuvre avec succès dans des domaines sensibles comme (la santé, l'IoT et les télécommunications) Zuo *et al.* (2022); Li *et al.* (2024). Un exemple significatif est l'utilisation de l'apprentissage fédéré par Google pour permettre d'optimiser les prévisions de claviers virtuels tout en maintenant la confidentialité des différents utilisateurs Yang *et al.* (2018). Malgré ces différents avantages, l'apprentissage fédéré demeure vulnérable à plusieurs menaces de sécurité et de confidentialité. Les attaques de type d'empoisonnement de modèle (model poisoning), l'inférence à partir des gradients et la récupération des données peuvent mettre en péril la confidentialité des données et l'intégrité du modèle global Bonawitz *et al.* (2017). Les risques sont encore plus élevés dans un contexte post-quantique, où les ordinateurs quantiques mettent en péril les primitives cryptographiques traditionnellement utilisées pour sécuriser les communications en apprentissage fédéré. Des algorithmes quantiques tels que l'algorithme de Shor permettent en effet de casser efficacement des schémas cryptographiques largement déployés comme RSA ou les courbes elliptiques mettant ainsi en péril la confidentialité et l'intégrité des échanges client-serveur Paillier (1999); Shor (1999). La cryptographie post-quantique (PQC) a été identifiée comme une solution prometteuse pour assurer une sécurité à long terme contre des adversaires dotés de capacités quantiques. Cependant,l'intégration de mécanismes post-quantiques dans les différents systèmes d'apprentissage fédérés est aujourd'hui confrontée à plusieurs défis. Les algorithmes post-quantiques

entrainants généralement un coût computationnel et communicationnel plus élevé, qui sont susceptibles de compromettre les performances de l'apprentissage, en particulier pour des tâches intensives comme la détsection d'objets Zhang *et al.* (2021). De plus, il est aussi essentiel de sécuriser les différents canaux de communication pour assurer la confidentialité en apprentissage fédéré, car le serveur central a toujours la possibilité d'utiliser les mises à jour locales pour accéder à des informations sensibles.

Pour répondre à ces enjeux, nous adoptons une approche de sécurité hybride comprenant des mécanismes cryptographiques post-quantiques pour protéger les communications et des mécanismes de protsection au niveau des gradients. Ainsi l'adoptons d'un cadre d'apprentissage fédéré standard repose sur un serveur central honnête, mais curieux, c'est a dire qu'il exécute correctement le mécanisme d'apprentissage et d'agrégation, tout en étant susceptible de tenter d'inférer des informations qui sont sensibles à partir des différentes mises à jour du modèle reçues, notamment au moyen d'attaques d'inférence ou de reconstruction. En parallèle, on estime que des adversaires externes ont la capacité d'observer, d'intercepter, de modifier ou d'injecter des messages sur les canaux de communication. Ce modèle de menace correspond aux hypothèses fréquemment admises dans la littérature sur l'apprentissage fédéré sécurisé Zuo *et al.* (2022); Bonawitz *et al.* (2017); Zhang *et al.* (2021); Beullens *et al.* (2021).

Pour garantir la confidentialité, un mécanisme d'encapsulation de clés post-quantiques est mis en place, basé sur des problèmes difficiles sur les réseaux euclidiens tels que CRYSTALS-Kyber, pour évaluer la résistance des clés aux attaques quantiques. Les signatures numériques post-quantiques, en particulier CRYSTALS-Dilithium sont utilisées pour authentifier les entités conformément aux recommandations actuelles du NIST et aux lignes directrices pour la migration vers la cryptographie post-quantique Chen *et al.* (2016, 2017). Afin d'empêcher le serveur d'accéder aux mises à jours individuels du modèle, le processus d'apprentissage est renforcé par un mécanisme d'agrégation sécurisée basé sur un masquage pairwise des

gradients dans une topologie de graphe de voisinage (Neighbors graph), garantissant que les gradients individuels ne peuvent être isolés lors de l'agrégation Bonawitz *et al.* (2017); Yang *et al.* (2022). Cette protsection est également consolidée par l'ajout d'un bruit gaussien déterministe utilisé comme un masque temporaire et annulable pour briser les corrélations statistiques exploitables avant l'agrégation, tout en assurant que le serveur n'a jamais accès aux gradients individuels en clair (plaintext), même de façon transitoire Paillier (1999); Shor (1999).

Sur la base de ce recherche, les questions suivantes sont étudiées :

RQ1. Est-ce qu'un mécanisme d'agrégation sécurisée incluant des mécanismes cryptographiques post-quantiques peut garantir une protsection efficace des gradients en apprentissage fédéré ?

RQ2. Quel est l'effet du mécanisme proposé sur la résistance aux attaques d'inférence et de poisoning par rapport à un apprentissage fédéré non protégé ?

RQ3. Est-ce que le mécanisme proposé maintient son efficacité lorsqu'il est utilisé dans les différentes architectures de détsection d'objets ?

Dans le but de répondre à ces questions, une évaluation expérimentale a été effectuée sur des tâches de détsection d'objets, en utilisant diverses architectures issues des familles YOLO et SSD. L'objectif est d'évaluer l'impact des mécanismes de sécurité proposés sur la confidentialité des gradients ainsi que sur les performances du modèle global, sans prétendre fournir des preuves cryptographiques formelles ni réaliser une comparaison exhaustive entre les algorithmes post-quantiques, conformément aux limites méthodologiques adoptées dans cette étude Tasnim & Qi (2023); Arya *et al.* (2021); Redmon *et al.* (2015).

Le reste de ce document est organisé de manière suivante. En effet, le contexte général et le modèle de menace sont présentés dans la section II. Ainsi, la section III décrit la méthode suggérée pour l'apprentissage fédéré sécurisé. Les travaux connexes sont examinés dans la section IV. Ensuite, la section V détaille le cadre expérimental, tandis que la section VI présente les résultats obtenus. Pour finir, la section VII discute ces résultats ainsi que leurs limites et la section VIII conclut l'étude en ouvrant des perspectives de recherche futures.

CHAPITRE I

REVUE DE LA LITTÉRATURE

L'apprentissage fédéré et la cryptographie post-quantique suscitent un intérêt croissant. Les inquiétudes portant sur la confidentialité des données et la sécurité des systèmes d'apprentissage automatique distribués augmentent. Alors, l'apprentissage fédéré évite le partage direct des données brutes entre participants, mais ce dernier reste vulnérable à de nombreuses attaques exploitant les paramètres échangés, surtout les gradients tels que démontrés par plusieurs travaux que cela peut permettre à un attaquant de reconstituer des informations sensibles sur les données d'origine Bonawitz *et al.* (2017); Redmon *et al.* (2015).

La littérature actuelle met en lumière plusieurs catégories de menaces pour l'apprentissage fédéré, incluant les attaques par inversion de gradients, l'empoisonnement de modèles et les attaques d'inférence sur les distributions locales de données. Ces failles ont provoqué l'apparition de mécanismes d'agrégation sécurisée, de techniques de masquage et de solutions cryptographiques pour améliorer la confidentialité et l'intégrité des mises à jour locales Bonawitz *et al.* (2017); Yang *et al.* (2022); Ultralytics (2024). Les attaques par fuite et la reconstruction de gradients tel que Deep Leakage from Gradients (DLG), surligne qu'un serveur ou un adversaire ayant accès aux gradients exacts peuvent reconstruire des données visuelles sensibles particulièrement lorsque les batchs sont de petites tailles Redmon *et al.* (2015). En outre, la montée en puissance de l'informatique quantique remet en question la sécurité des schémas cryptographiques traditionnels utilisés pour sécuriser les communications et l'authentification dans les systèmes fédérés. Les travaux récents illustre qu'un adversaire quantique exploitant l'algorithme de Shor peut casser des mécanismes d'établissement de clés tels que RSA et les courbes elliptiques (ECC) en temps polynomial compromettant ainsi la confidentialité à long terme des échanges FL ce qui est un risque majeur pour les systèmes

"harvest-now, decrypt-later" Shor (1999). Dans le domaine de la vision par ordinateur fédérée, l'intégration de modèles profonds de détection d'objets, notamment YOLO illustre une forte efficacité de temps réel et également une sensibilité accrue aux fuites d'information via les gradients ainsi qu'aux perturbations introduites par les mécanismes de sécurité Arya *et al.* (2021); Redmon *et al.* (2015); Sawatzky *et al.* (2019). Cette revue de la littérature vise à présenter de manière structurée les fondements théoriques de l'apprentissage fédéré, ses principaux algorithmes, les défis liés aux données non-IID, ainsi que les applications en détection d'objets pour des systèmes de vision profonde. Elle établit le cadre conceptuel nécessaire à la compréhension des travaux proposés dans cette recherche. De plus, cette dernière justifie les choix méthodologiques et cryptographiques adoptés par la suite, notamment l'intégration de mécanismes d'agrégation sécurisée, de masquage pairwise ainsi que la cryptographie post-quantique normalisée dans les pipelines FL Bonawitz *et al.* (2017); Shor (1999); Yang *et al.* (2022); Sawatzky *et al.* (2019).

1.1 DÉFINITION ET ARCHITECTURE DE L'APPRENTISSAGE FÉDÉRÉ

L'apprentissage fédéré (Fédérational Learning, FL) est un paradigme d'apprentissage automatique distribué dans lequel plusieurs entités collaborent à l'entraînement d'un modèle global sans jamais partager leurs données brutes. Introduit pour répondre aux enjeux de confidentialité et de conformité réglementaire, le FL repose sur le principe selon lequel les données restent stockées localement chez chaque participant, tandis que seuls les paramètres du modèle tels que les poids ou gradients sont échangés avec un serveur central chargé de l'agrégation Bonawitz *et al.* (2017); Beullens *et al.* (2021). L'architecture standard de l'apprentissage fédéré est basée sur un schéma client-serveur. Le serveur crée un modèle global qu'il partage avec les clients participants. Chaque client intègre localement ce modèle dans son propre ensemble de données, puis envoie les mises à jour calculées au serveur.

Celui-ci agrège les contributions reçues afin de produire un nouveau modèle global qui sera redistribué ensuite aux clients lors du round suivant. On continue cette boucle itérative jusqu'à ce qu'elle se converge Bonawitz *et al.* (2017). Bien que cette architecture limite l'exposition directe des données, elle introduit de nouveaux défis liés à la sécurité des échanges, à la fiabilité des clients et à la confidentialité des informations pouvant être inférées à partir des paramètres partagés Yang *et al.* (2022); Heik *et al.* (2025).

1.2 ALGORITHME FEDAVG ET VARIANTES

FedAvg (Federated Averaging) est la méthode d'agrégation la plus couramment utilisée en apprentissage fédéré. Il est basé sur l'agrégation des mises à jour locales sous la forme d'une moyenne pondérée en fonction de la taille des jeux de données détenus par les clients, selon les recommandations établies dans les travaux fondateurs du FL Kaur & Singh (2022). Malgré sa simplicité et son efficacité, FedAvg a des limitations majeures, en particulier lorsqu'il y a des données hétérogènes ou des comportements clients intermittents, peu fiables ou malveillants. Ces limites ont conduit au développement de variantes telles que FedProx, FedNova et FedOpt, visant à améliorer la stabilité et la convergence du modèle global dans des environnements distribués réalistes en particulier sous des distributions de données non-IID Yang *et al.* (2022); Sawatzky *et al.* (2019). Toutefois, ces approches d'agrégation améliorées ne traitent pas explicitement les problématiques de confidentialité et de sécurité des gradients, qui restent vulnérables aux attaques de fuite d'information, d'inversion et d'inférence, ainsi qu'aux menaces de poisoning et backdoor FL qui sont introduites par des clients compromit Zuo *et al.* (2022); Heik *et al.* (2025).

1.3 PROBLÉMATIQUE DES DONNÉES NON-IID

Le FL est confronté à un défi majeur en raison de la nonindépendance et de l'identité des distributions de données locales. Les données des clients sont fréquemment issues de distributions hétérogènes influencées par des facteurs contextuels variés dans des scénarios concrets. Cette hétérogénéité statistique peut ralentir la convergence et dégrader les performances du modèle global Zhang *et al.* (2021); Abdi (2007). Les données non-IID augmentent aussi les risques de fuite d'informations lors du partage des gradients. Ainsi, les attaques comme Deep Leakage from Gradients (DLG) démontrent comment un adversaire qui possède les gradients exacts peut reconstruire des données privées et révéler des caractéristiques sensibles des distributions locales. L'algorithme FedAvg, bien qu'efficace et simple, présente des limites lorsque les données sont fortement hétérogènes ou Non-IID. Ce qui motive le développement de variantes en améliorant la convergence, mais pas la confidentialité des gradients Kaur & Singh (2022). Cette problématique souligne l'importance d'implémenter des mécanismes de protection appropriés, notamment l'agrégation sécurisée et la dissimulation des mises à jour locales, pour limiter la fuite d'informations au-delà de l'optimisation de la convergence Bonawitz *et al.* (2017).

1.4 APPLICATIONS EN VISION PAR ORDINATEUR

L'apprentissage fédéré a été mis en œuvre dans divers domaines de la vision par ordinateur tels que la classification d'images et la détection d'objets. Ces applications sont spécialement adaptées à des contextes où les données visuelles sont sensibles et distribuées, comme la vidéo surveillance, les véhicules autonomes et les systèmes edge/IoT Abdi (2007). Les modèles de détection d'objets de la famille YOLO introduits dans les architectures one-stage de vision profonde se distinguent par leur efficacité en temps réel, mais leur intégration dans un FL soulève des défis liés à la taille des mises à jour et à l'exposition structurelle

des gradients Kaur & Singh (2022); Abdi (2007). Ces modèles sont exposés à des attaques par inversion ou reconstruction de gradients comme le montrent les travaux sur la fuite de gradients en FL Heik *et al.* (2025). Ainsi, même si des variantes de FedAvg améliorent la convergence, la confidentialité et l'intégrité des gradients dans un FL appliqué à la vision, mais elles ne sont pas garanties sans l'ajout de mécanismes cryptographiques renforcés et de protections complémentaires Bonawitz *et al.* (2017); Chen *et al.* (2016).

1.5 ATTAQUES CONTRE L'APPRENTISSAGE FÉDÉRÉ

L'apprentissage fédéré, bien qu'il limite le partage explicite des données brutes, mais il n'élimine pas les risques de fuite d'informations sensibles. De nombreux travaux ont démontré que les paramètres échangés durant le processus d'apprentissage en particulier les gradients et les poids intermédiaires peuvent être exploités par un adversaire pour inférer des informations privées sur les données locales des clients. Dans cette section, il sera question d'examiner les principales catégories d'attaques documentées dans la littérature qui ciblent l'apprentissage fédéré en mettant l'accent sur celles qui exploitent les gradients et les mises à jour locales Heik *et al.* (2025); Abdi (2007).

1.5.1 DEEP LEAKAGE FROM GRADIENTS (DLG)

L'attaque Deep Leakage from Gradients (DLG) est l'un des premiers exemples concrets montrant que les gradients partagés dans un contexte d'apprentissage collaboratif peuvent révéler directement les données d'entraînement. Zhu, Liu et Han ont proposé cette attaque qui vise à reconstruire les données d'entrée et leurs labels en optimisant une entrée fictive avec des gradients similaires à ceux observés lors d'un round d'apprentissage. DLG suppose qu'un adversaire peut accéder aux gradients exacts transmis par un client au serveur. En résolvant un problème d'optimisation inverse, l'attaque peut reconstruire des images d'entraînement

avec une fidélité élevée, surtout lorsque les lots sont petits. L'attaque a révélé une vulnérabilité structurelle du FL en montrant que le simple partage des données brutes ne garantit pas la confidentialité. Dans la littérature, Secure Agrégation est une solution utilisée pour masquer les gradients individuels avant l'agrégation, mais elle n'est pas suffisante pour empêcher totalement la fuite d'information restante lors des attaques inverses Bonawitz *et al.* (2017); Heik *et al.* (2025); Frankel *et al.* (2003).

1.5.2 IDLG (IMPROVED DLG)

Pour optimiser l'efficacité de DLG (Zhao et al), ont avancé une méthode connue sous le nom d'Improved Dee Leakage from Gradients (iDLG). Cette attaque exploite les propriétés supplémentaires des différents gradients, en particulier la structure des gradients qui sont associés à la dernière couche du réseau, pour inférer directement les labels sans l'optimisation complète. (iDLG) permets d'améliorer de manière significative la vitesse et la stabilité de la reconstruction par rapport à DLG, tout en limitant le besoin des contraintes supplémentaires. Cette amélioration rend les attaques par fuite de gradients encore plus réalistes dans des scénarios d'apprentissage fédéré, en particulier lorsque les gradients sont transmis sans protection cryptographique ou sans mécanisme de masquage Heik *et al.* (2025).

1.5.3 ATTAQUES D'INVERSION DE MODÈLE

Les attaques d'inversion de modèle ont pour but de reconstruire les caractéristiques des données d'entraînement en se basant sur les paramètres finaux ou intermédiaires d'un modèle entraîné Shor (1999). Contrairement à DLG, ces attaques ne requièrent pas l'accès aux gradients détaillés d'un round spécifique, mais ils utilisent les poids du modèle global ou ses sorties Paillier (1999). Dans un environnement fédéré, elles peuvent mettre en lumière

des attributs sensibles sur les données locales des clients, surtout lorsque le modèle est surdimensionné ou appliqué à des données fortement non-IID Bonawitz *et al.* (2017); Ultralytics (2024). Les modèles de vision profonde sont particulièrement susceptibles d'être attaqués par ce type d'attaque, car ils ont la capacité de mémoriser des structures complexes qui simplifient l'inversion Tasnim & Qi (2023); Arya *et al.* (2021)

1.5.4 ATTAQUES D'INFÉRENCE D'APPARTENANCE

L'objectif des attaques d'inférence d'appartenance est de déterminer si un échantillon donné a été inclus ou non dans le jeu de données d'entraînement d'un modèle. Au départ, ces attaques ont été étudiées dans un contexte centralisé, mais elles sont désormais utilisées pour inférer la participation d'un client ou d'un échantillon spécifique lors de l'apprentissage fédéré en exploitant les mises à jour successives du modèle global. Dans le FL, un adversaire a la capacité d'analyser les variations des paramètres du modèle sur plusieurs rounds pour mettre en évidence des liens avec certaines données privées. Ces attaques représentent un risque majeur pour la confidentialité en particulier dans des domaines sensibles comme la santé ou la biométrie, où la simple appartenance à un jeu de données peut constituer une information critique Bonawitz *et al.* (2017); Frankel *et al.* (2003).

1.5.5 ATTAQUES PAR CORRÉLATION INTER-ROUNDS

Les attaques par corrélation inter-rounds exploitent la nature itérative de l'apprentissage fédéré. En observant les gradients ou les poids agrégés sur plusieurs rounds successifs, un adversaire peut établir des corrélations temporelles permettant d'isoler la contribution d'un client spécifique ou d'inférer des propriétés de ses données locales. Ces attaquants sont extrêmement efficaces lorsque le nombre de clients participants est limité que les données sont

fortement non-IID, ou lorsque les mêmes clients participent souvent aux rounds successifs. La littérature montre qu'en l'absence de mécanismes de randomisation ou de protections supplémentaire, l'accumulation d'informations au fil des rounds peut conduire à des fuites de données substantielles Bonawitz *et al.* (2017); Frankel *et al.* (2003).

1.6 SOLUTIONS EXISTANTES DE SÉCURISATION

Les attaques d'inversion du modèle ont pour but de reconstruire les caractéristiques des données d'entraînement en se basant sur les paramètres finaux ou intermédiaires d'un modèle entraîné Zhang *et al.* (2021); Frankel *et al.* (2003). Contrairement à DLG, ces attaques ne nécessitent pas d'accéder aux gradients détaillés d'un round spécifique, mais utilisent les poids du modèle global ou ses sorties Heik *et al.* (2025). Dans un environnement fédéré, elles peuvent exposer des caractéristiques sensibles dans les données locales des clients en particulier lorsque le modèle est surdimensionné ou entraîné sur des données fortement non-IID. Ce type d'attaque est particulièrement risqué pour les modèles de vision profonde, car ils peuvent mémoriser des structures complexes qui facilitent l'inversion Zhang *et al.* (2021); Ultralytics (2024); Frankel *et al.* (2003).

1.6.1 TLS ET AES

La première approche de sécurisation dans les systèmes d'apprentissage fédéré consiste à chiffrer les canaux de communication entre les clients et le serveur, généralement à l'aide de protocoles standards tels que TLS, combinés à des algorithmes de chiffrement symétrique comme AES Gurung *et al.* (2023). Ce mécanisme permet de se protéger contre les attaques passives, notamment en écoutant et en interceptant les communications sur le réseau Yang *et al.* (2018); Liu *et al.* (2024). Cependant, seules les données en transit sont protégées par le chiffrement du canal. Une fois les gradients déchiffrés du côté serveur, ils restent visibles en

temps réel en exposant le système aux attaques internes ou aux serveurs considérés comme "honnêtes, mais-sécurieux". En conséquence, même si c'est indispensable, le chiffrement du canal ne suffit pas à garantir la confidentialité des gradients dans un environnement d'apprentissage fédéré Bonawitz *et al.* (2017).

1.6.2 AGRÉGATION SÉCURISÉ

Le protocole de l'agrégation Sécurisé a pour objectif d'empêcher le serveur d'accéder aux gradients individuels tout en lui offrant la capacité de calculer leur agrégation globale. L'approche suggéré par Bonawitz et al. est basé sur un mécanisme de masquage pairwise, où chaque client masque ses gradients en partageant des valeurs aléatoires avec d'autres clients. Lors de l'agrégation, les masques s'annulent mathématiquement, ce qui permet au serveur d'observer uniquement la somme globale des gradients. Ce mécanisme offre une forte protection contre les attaques par fuite de gradients, même en présence d'un serveur honnête, mais curieux et il a démontré sa faisabilité à grande échelle dans des environnements réels. Néanmoins, les mécanismes de sécure agrégation introduisent une complexité supplémentaire, notamment en matière de gestion des clients défaillants, de surcharge de communication et de synchronisation. En outre, ils ne prennent pas en charge de manière intrinsèque les menaces liées à l'authentification des mises à jour ou aux attaques actives comme l'empoisonnement de modèles ou l'injection de gradients malveillants Bonawitz *et al.* (2017); Zhang *et al.* (2021); Frankel *et al.* (2003).

1.6.3 CONFIDENTIALITÉ DIFFÉRENTIELLE

La confidentialité différentielle (Differential Privacy, DP) constitue une approche probabiliste qui vise à limiter la quantité d'informations qu'un adversaire peut inférer sur la donnée

individuelle à partir des différentes sorties d'un algorithme. Dans le cadre de l'apprentissage fédéré, la (DP) est couramment mise en œuvre en ajoutant un bruit aléatoire qui est adapté aux gradients ou aux paramètres du modèle avant de les envoyer au serveur. Cette méthode offre des garanties formelles de confidentialité et protège contre certaines attaques d'inférence, y compris les attaques d'appartenance. Cependant, la présence de bruit a un impact direct sur la précision et la convergence du modèle, surtout pour des tâches complexes telles que la vision par ordinateur. Le choix du compromis entre confidentialité et performance reste délicat et une confidentialité élevée peut causer une détérioration significative des performances du modèle Yang *et al.* (2022); Abdi (2007); Frankel *et al.* (2003).

1.6.4 CHIFFREMENT HOMOMORPHE

Le chiffrement homomorphe permet d'effectuer des opérations arithmétiques directement sur des données chiffrées, sans nécessiter leur déchiffrement préalable. Appliqué à l'apprentissage fédéré, ce principe permet au serveur de faire l'agrégation des gradients chiffrés sans pouvoir accéder aux valeurs en clair et offrant ainsi un niveau élever de confidentialité sur le plan théorique. Cependant, les schémas de chiffrement homomorphe restent extrêmement coûteux en termes de calcul et de latence. Leur intégration dans des systèmes d'apprentissage fédéré à grande échelle est souvent impraticable, en particulier pour des différents modèles profonds comportant un grand nombre de paramètres. En conséquence, bien qu'il soit prometteur sur le plan théorique, cette approche est rarement utilisée seule dans des systèmes FL opérationnels Zhang *et al.* (2021); Frankel *et al.* (2003).

1.6.5 CALCUL MULTIPARTITE SÉCURISÉ (MPC)

Le Calcul multipartite sécurisé (MPC) rassemble un ensemble de protocoles qui permettent à plusieurs parties de calculer ensemble une fonction sur leurs entrées privées, sans jamais divulguer ces entrées aux autres participants. Dans le cadre de l'apprentissage fédéré, la MPC peut être employée pour assurer une agrégation sécurisée des gradients en empêchant le serveur d'accéder aux contributions individuelles des clients. Même si la MPC garantit la confidentialité de manière solide, elle rencontre des limitations importantes en termes de complexité protocolaire, de surcharge de communication et de dépendance à un nombre élevé de participants actifs. Ces contraintes rendent son déploiement complexe dans des environnements à grande échelle, avec des clients hétérogènes et intermittents, typiques des systèmes d'apprentissage fédéré réel Yang *et al.* (2022); Abdi (2007); Frankel *et al.* (2003).

La littérature démontre que les solutions actuelles de sécurisation de l'apprentissage fédéré présentent toutes des garanties partielles. Le chiffrement du canal protège les communications, mais pas les gradients en mémoire serveur, la confidentialité différentielle introduit un compromis délicat entre la protection et la performance, tandis que le chiffrement homomorphe et la MPC restent coûteux et peu scalables. Les protocoles de sécure agrégation constituent actuellement l'une des solutions les plus adoptées pour protéger les gradients, mais ils nécessitent d'être combinés à d'autres mécanismes pour répondre aux menaces complètes identifiées dans les sections précédentes Zhang *et al.* (2021); Bonawitz *et al.* (2017).

1.7 LIMITES DES SOLUTIONS EXISTANTES

Malgré les avancées significatives proposées par les mécanismes de sécurisation de l'apprentissage fédéré, la littérature met en évidence plusieurs limites structurelles qui restreignent leur efficacité, en particulier lorsqu'ils sont appliqués à des modèles de vision

profonde et à des scénarios de données non-IID. Cette section examine de manière critique les principales contraintes associées aux solutions existantes, ce qui explique pourquoi il est essentiel d'adopter un mécanisme de sécurité hybride et post-quantique tel que celui proposé dans ce travail Zhang *et al.* (2021).

1.7.1 FUITE RÉSIDUELLE DES GRADIENTS

Malgré l'utilisation de mécanismes tels que l'agrégation sécurisée ou la confidentialité différentielle pour réduire l'exposition directe des gradients, plusieurs études démontrent que des fuites d'information persistent dans certaines conditions. Les attaques de reconstruction de gradients, notament Deep Leakage from Gradients (DLG) et ses variantes restent partiellement efficaces lorsque les gradients sont protégés de manière insuffisante, ou lorsqu'il est possible d'exploiter des corrélations inter-rounds. Ces fuites résiduelles jouent un rôle crucial dans les modèles de vision profonde, car les gradients restent fortement en corrélation avec les structures spatiales des données d'entrée. Même après l'ajout de bruit ou de masques incomplets, un serveur honnête, mais curieux peut exploiter des signaux statistiques pour inférer des informations sensibles Zhang *et al.* (2021); Heik *et al.* (2025).

1.7.2 COÛT COMPUTATIONNEL ET COMMUNICATION

Les solutions de sécurisation actuelles entraînent un surcoût important en termes de calcul et de communication. Les protocoles d'agrégation de sécurité nécessitent de multiples échanges de clés entre les clients, tandis que les approches basées sur la confidentialité différentielle exigent des opérations supplémentaires pour ajuster le bruit. Les approches plus lourdes telles que le chiffrement homomorphe ou le multiparty computation (MPC) amplifient ces coûts de manière importante les rendant difficilement compatibles avec des scénarios

réels à grande échelle et avec des modèles profonds comportant des millions de paramètres Bonawitz *et al.* (2017); Zhang *et al.* (2021); Das & Das (2025).

1.7.3 DÉGRADATION DE LA CONVERGENCE

La littérature met également en évidence une détérioration notable des performances et de la vitesse de convergence lorsque des mécanismes de protection sont appliqués de manière agressive. L'ajout de bruit dans le contexte de la confidentialité différentielle peut entraîner une diminution de la précision et de l'instabilité de l'entraînement, en particulier lorsque les données sont hautement non-IID Zhang *et al.* (2021). Dans le cas des modèles de vision par ordinateur, comme YOLO, cette dégradation est encore plus marquée en raison de la sensibilité des détecteurs d'objets aux perturbations des gradients, affectant directement les métriques de détection telles que la précision, le rappel et la mAP Tasnim & Qi (2023); Ultralytics (2024); Das & Das (2025).

1.7.4 ABSENCE DE RÉSISTANCE POST-QUANTIQUE

Un autre aspect critique des solutions existantes est leur dépendance aux primitives cryptographiques classiques. La majorité des implémentations d'apprentissage fédéré sécurisées utilisent toujours RSA, Diffie-Hellman ou les courbes elliptiques (ECC) pour établir des clés et s'authentifier. Les schémas sont vulnérables aux ordinateurs quantiques, en particulier via l'algorithme de Shor, ce qui met en péril la sécurité à long terme. En l'absence d'intégration explicite de mécanismes post-quantiques, les protocoles actuels ne sont pas sûrs de préserver à long terme les communications et les mises à jour locales face à une menace quantique émergente Paillier (1999).

1.7.5 INCOMPATIBILITÉ AVEC MODÈLES PROFONDS

Enfin, de nombreuses solutions de protection ont été conçues et évaluées principalement sur des modèles simples ou de petites tailles. Lorsqu'elles sont utilisées dans des architectures de vision profonde comme YOLO, elles rencontrent des problèmes de scalabilité, de latence excessive et de complexité mémoire Zhang *et al.* (2021). Les gradients volumineux, la fréquence élevée des échanges et la sensibilité des modèles de détection d'objets aux perturbations rendent ces solutions difficilement exploitables en pratique sans une conception spécifique tenant compte des contraintes propres à la vision par ordinateur fédéré Tasnim & Qi (2023); Arya *et al.* (2021). Cette analyse met en évidence que les solutions existantes, bien qu'efficaces sur certains aspects, ils présentent des limites majeures en matière de fuite résiduelle, de coût, de convergence, de résistance post-quantique et de compatibilité avec les modèles de vision profonde. Ces constats justifient la mise en place d'une approche hybride combinant masquage pairwise, cryptographie post-quantique et chiffrement symétrique efficace, comme celle suggérée dans cette recherche Zhang *et al.* (2021).

1.8 CRYPTOGRAPHIE POST-QUANTIQUE

L'informatique quantique évolue rapidement et remet en question les bases de la cryptographie classique qui est utilisée dans la plupart des systèmes distribués actuels, y compris l'apprentissage fédéré. Les protocoles de sécurisation traditionnels ne garantissent plus une sécurité à long terme face à des adversaires dotés de capacités quantiques. Cette section présente les menaces quantiques pesant sur la cryptographie classique, introduit les schémas KEM post-quantiques et met en évidence le rôle central de CRYSTALS-Kyber dans la conception de protocoles d'apprentissage fédéré résistants à la quantique Shor (1999); Gorbenko & Kandii (2025).

1.8.1 MENACE QUANTIQUE

Les ordinateurs classiques sont principalement confrontés à des problèmes mathématiques difficiles, tels que la factorisation de grands entiers (RSA) ou le logarithme discret sur les courbes elliptiques (ECC), ce qui rend la cryptographie asymétrique classique essentielle. Cependant, l'algorithme quantique de Shor démontre qu'un ordinateur quantique suffisamment puissant peut régler ces problèmes dans des temps polynomiaux, ce qui élimine les garanties de sécurité offertes par RSA, Diffie-Hellman et ECC. Cette vulnérabilité est d'une importance capitale dans un contexte d'apprentissage fédéré. Les canaux de communication client-serveur font souvent usage de ces primitives pour échanger des clés et authentifier. Un adversaire quantique pourrait intercepter et décoder de façon rétroactive les communications, ce qui mettrait en péril la confidentialité des gradients, des mises à jour de modèle et des poids globaux échangés. Cette menace, dite Harvest now, decrypt later, constitue un risque majeur pour les systèmes d'apprentissage fédérés déployés sur de longues périodes Shor (1999); Wang & Bovik (2009).

1.8.2 SCHÉMAS KEM

Pour répondre à ces menaces, la communauté scientifique s'est tournée vers la cryptographie post-quantique (PQC), qui vise à concevoir des primitives cryptographiques résistant à la fois aux attaques classiques et quantiques. Dans le cadre des différentes communications sécurisées, les mécanismes d'encapsulation des clés (Key Encapsulation Mechanisms KEM) jouent un rôle central, en permettant de faire l'établissement des clés sécrète partagés sans vulnérabilité connue face aux différents algorithmes post-quantiques. Les processus de standardisation du (NIST) ont identifié plusieurs familles de schémas KEM post-quantiques, notamment les approches basées sur les réseaux euclidiens (lattice-based), les codes de correcteurs d'erreurs et les isogénies de courbes elliptiques. Parmi celles-ci, les schémas lattice-based

se distinguent par leur efficacité, leur maturité théorique et leur adaptabilité aux systèmes distribués. Dans l'apprentissage fédéré, l'intégration de KEM post-quantique permet d'établir des clés symétriques robustes entre clients et serveur, servant ensuite au chiffrement efficace des gradients et des modèles à l'aide de primitifs symétriques légers Li *et al.* (2024); Gurung *et al.* (2023).

1.8.3 KYBER

Le NIST a choisi CRYSTALS-Kyber comme standard post-quantique pour l'établissement de clés. Il est basé sur la difficulté du problème du Module Learning With Errors (Module-LWE), qui est réputé pour sa capacité à résister aux attaques quantiques connues. Kyber se décline en plusieurs niveaux de sécurité, dont Kyber512, offrant un compromis adapté entre sécurité et performance Gorbenko & Kandii (2025).

Plusieurs études ont montré que Kyber présente un coût computationnel modéré, des tailles de clés et de ciphertexts raisonnables, et une implémentation efficace sur des environnements contraints, ce qui le rend particulièrement adapté aux scénarios distribués et à l'apprentissage fédéré. Kyber512 est utilisé dans le but d'établir des clés partagées entre le serveur et les clients, ainsi qu'entre pairs dans le cadre du masquage pairwise. L'utilisation de Kyber permet ainsi de garantir la confidentialité des échanges, même face à un adversaire quantique, tout en conservant une efficacité compatible avec les besoins des modèles de vision profonde Zhang *et al.* (2021).

1.8.4 COMPARAISON AVEC RSA ET ECC

Kyber est plus résistant aux attaques quantiques que RSA et ECC, contrairement aux schémas classiques qui reposent sur des hypothèses devenues invalides dans un contexte

quantique. Bien que les clés et ciphertexts de Kyber soient plus volumineux que ceux d'ECC, leur impact reste acceptable dans des architectures FL moderne, en particulier lorsque Kyber est utilisé uniquement pour l'échange initial de clés et non pour le chiffrement massif des données. Par ailleurs, le chiffrement symétrique de Kyber via HKDF (par exemple AES-CBC ou AES-GCM) offre à la fois sécurité post-quantique et efficacité pratique. Cette approche hybride est actuellement perçue comme une stratégie réaliste pour passer de systèmes sécurisés qui ne supportent pas la quantique, notament dans des applications sensibles telles que la vision par ordinateur fédéré Zhang *et al.* (2021); Kaur & Singh (2022).

La cryptographie post-quantique joue un rôle crucial dans la sécurité à long terme de l'apprentissage fédéré. Les limites intrinsèques de RSA et d'ECC face aux menaces quantiques rendent indispensable l'adoption de schémas KEM post-quantiques. CRYSTALS-Kyber, en particulier Kyber512, s'impose comme une solution adaptée pour sécuriser les échanges dans les systèmes FL modernes, tout en maintenant une efficacité compatible avec les contraintes computationnelles et communicationnelles des modèles de vision profonde Zhang *et al.* (2021); Abdi (2007).

1.8.5 MODÈLES DE DÉTECTION D'OBJETS

La détection d'objets est une étape cruciale de la vision par ordinateur, car elle permet de localiser et de classer simultanément plusieurs objets dans une image ou une séquence vidéo. Contrairement à la classification habituelle des images, la détection implique une estimation précise des boîtes englobantes (bounding boxes) et l'identification des classes associées, ce qui implique des modèles profonds plus complexes et coûteux sur le plan computationnel. Dans un contexte d'apprentissage fédéré, ces caractéristiques posent des défis spécifiques en ce qui concerne la confidentialité, la communication et la convergence Tasnim & Qi (2023); Arya *et al.* (2021).

1.8.6 RÉSEAUX CONVOLUTIFS POUR LA DÉTECTION

Les réseaux de neurones convolutifs (CNN) sont essentiels pour les modèles modernes de détection d'objets. Ils autorisent l'extraction automatique de caractéristiques hiérarchiques à partir des données visuelles, de motifs locaux simples à des représentations sémantiques complexes. Les premières méthodes de détection étaient basées sur des architectures en plusieurs étapes, comme R-CNN, Fast R-CNN et Faster R-CNN, qui séparent l'extraction de caractéristiques, la création de propositions de régions et la classification finale. Malgré leur précision, ces architectures multiétapes souffrent d'une latence élevée et d'une complexité computationnelle importante, les rendant ainsi moins adaptées aux scénarios temps réel et aux environnements distribués contraints. Ces limitations ont engendré la création de modèles one-stage, capables de détecter en une seule passe du réseau, ouvrant ainsi la voie à des architectures plus efficaces telles que YOLO et SSD Yang *et al.* (2022); Redmon *et al.* (2015).

1.8.7 PRÉSENTATION DE YOLO

Les modèles YOLO (You Only Look Once) ont révolutionné la détection d'objets en adoptant une approche unifiée qui traite la détection comme un problème de régression directe à partir de l'image complète. Contrairement aux méthodes basées sur des régions, YOLO fractionne l'image en une grille et il prévoit en parallèle les boîtes englobantes, les probabilités de classes et les scores de confiance. Cette approche permet d'atteindre des performances en temps réel tout en conservant une précision compétitive, ce qui explique son adoption massive dans des applications critiques telles que la vidéo surveillance, les véhicules autonomes et les systèmes embarqués. Les versions successives du YOLOv2 au YOLOv5 ont introduit des améliorations progressives en ce qui concerne l'architecture, les fonctions de perte et les stratégies d'entraînement Tasnim & Qi (2023); Arya *et al.* (2021).

1.8.8 YOLOV8 : ARCHITECTURE ET CARACTÉRISTIQUES

YOLOv8 représente une évolution récente de la famille YOLO, intégrant des optimisations architecturales destinées à améliorer à la fois la précision et l'efficacité computationnelle. Cette version adopte une architecture modulaire reposant sur des blocs convolutifs optimisés, des mécanismes d'agrégation de caractéristiques multiéchelles et une tête de détection améliorée, facilitant l'apprentissage de représentations plus robustes. YOLOv8 est décliné en plusieurs variantes (n, s, m, l, x), permettant un compromis flexible entre précision et coût de calcul. Cette adaptabilité en fait un candidat particulièrement pertinent pour l'apprentissage fédéré, où les clients peuvent disposer de capacités matérielles hétérogènes. Toutefois, la taille des gradients, la fréquence des mises à jour et la sensibilité des détecteurs aux perturbations soulèvent des défis majeurs lorsqu'il est entraîné dans un cadre fédéré sécurisé Tasnim & Qi (2023); Arya *et al.* (2021).

1.8.9 ENJEUX DE LA VISION PAR ORDINATEUR EN APPRENTISSAGE FÉDÉRÉ (FL)

L'intégration de modèles de détection d'objets dans un cadre d'apprentissage fédéré pose des défis spécifiques qui dépassent ceux rencontrés avec des modèles de classification standard. Tout d'abord, les gradients issus des modèles de vision profonde contiennent une information structurellement riche, fortement corrélée aux données d'entrée, ce qui les rend particulièrement vulnérables aux attaques de reconstruction telles que DLG et iDLG Paillier (1999); Redmon *et al.* (2015). Ensuite, il est fréquent que les données en vision par ordinateur ne soient pas non- IID, car chaque client peut observer des scènes, des classes d'objets ou des conditions d'éclairage très différentes. Les difficultés de convergence sont accentuées par cette hétérogénéité, ce qui peut sérieusement affecter les performances du modèle global si elle n'est pas correctement prise en compte. Les contraintes communicationnelles sont renforcées

par la taille importante des modèles et des gradients associés aux architectures de détection. Ces facteurs rendent nécessaire l'utilisation de mécanismes de sécurité efficaces, mais légers, capables de maintenir les gradients sans compromettre la convergence ni les performances de détection. Il est justifié par ces enjeux de choisir YOLOv8 comme modèle d'étude dans ce travail, tout en intégrant conjointement des mécanismes de protection cryptographiques et protocolaires adaptés à la vision par ordinateur fédéré Tasnim & Qi (2023); Arya *et al.* (2021).

1.9 CONCLUSION DU CHAPITRE

Ce chapitre a présenté une revue approfondie de la littérature relative à l'apprentissage fédéré en mettant en évidence à la fois ses principes fondamentaux, ses vulnérabilités et les limites des solutions de sécurisation existantes. L'apprentissage fédéré se positionne comme une approche clé pour préserver la confidentialité des données dans les systèmes distribués, mais la littérature montre clairement qu'il demeure exposé à des attaques sophistiquées visant les gradients, les mises à jour locales et le modèle global, en particulier dans des contextes de données non-IID et de modèles profonds Zuo *et al.* (2022); Paillier (1999); AbhishekV *et al.* (2022).

L'analyse des attaques majeures tel que Deep Leakage from Gradients, iDLG, les attaques d'inversion de modèle, d'inférence d'appartenance et les corrélations inter-rounds, a démontré que le simple fait de conserver les données localement ne suffit pas à garantir la confidentialité. Les gradients échangés contiennent une information structurelle exploitable, rendant indispensable l'intégration de mécanismes de protection spécifiques à l'apprentissage fédéré Paillier (1999); Sawatzky *et al.* (2019).

Les solutions existantes de sécurisation incluant le chiffrement des canaux, l'agrégation sécurisée, la confidentialité différentielle, le chiffrement homomorphe et la multiparty com-

putation offrent chacune des garanties partielles, mais au prix de compromis importants. La littérature met en évidence des limites récurrentes en termes de fuite résiduelle des gradients, de surcoûts computationnels et communicationnels, de dégradation de la convergence, d'incompatibilité avec les modèles de vision profonde et, surtout, d'absence de résistance face aux menaces post-quantiques Sawatzky *et al.* (2019); AbhishekV *et al.* (2022).

Par ailleurs, l'étude de la cryptographie post-quantique a montré que les primitives asymétriques classiques sur lesquelles reposent encore de nombreux systèmes FL sont insuffisantes à long terme. Les avancées en informatique quantique, notamment via l'algorithme de Shor, rendent nécessaire l'adoption de mécanismes cryptographiques résistants aux quantiques, parmi lesquels les schémas KEM lattice-based, et en particulier CRYSTALS-Kyber, apparaissent comme des solutions adaptées aux environnements distribués modernes Zhang *et al.* (2021); Beullens *et al.* (2021).

nfin, l'intégration de modèles de détection d'objets, tels que YOLOv8, dans un cadre d'apprentissage fédéré sécurisé soulève des défis supplémentaires. La taille des modèles, la richesse informationnelle des gradients et la sensibilité des performances aux perturbations rendent les approches de sécurisation classiques difficilement applicables sans une conception spécifique et optimisée Tasnim & Qi (2023); Arya *et al.* (2021).

Ainsi, cette revue de la littérature met en évidence un vide méthodologique clair. L'absence d'un mécanisme unifié capable de combiner efficacement la protection des gradients, l'authentification des mises à jour, une sécurité post-quantique et une compatibilité avec les modèles de vision profonde dans un contexte fédéré. Ces constats constituent la motivation principale du mécanisme proposé dans ce travail. Le chapitre suivant présente en détail la méthodologie adoptée, qui vise à combler ces lacunes en intégrant de manière cohérente la

cryptographie post-quantique, le masquage pairwise, le chiffrement symétrique et les signatures numériques au sein d'un cadre d'apprentissage fédéré appliqué à la détection d'objets.

CHAPITRE II

APPROCHE PROPOSÉE

Dans ce chapitre, il sera question de présenter la méthode proposée pour la conception et la mise en œuvre d'un mécanisme d'apprentissage fédéré sécurisé destiné à l'entraînement de modèles de détection d'objets. Le principal objectif est d'assurer la confidentialité des données locales, la préservation des gradients, l'intégrité des communications et la résistance aux attaques classiques et post-quantiques, tout en garantissant des performances d'apprentissage adéquates. La méthodologie repose sur une combinaison de mécanismes complémentaires, incluant l'apprentissage fédéré, la cryptographie post-quantique, le masquage pairwise des gradients, le chiffrement symétrique et les signatures numériques. Ce chapitre commence par définir les besoins de sécurité et le modèle de menace envisagé, puis passe en revue les hypothèses du système et les adversaires pris en compte.

2.1 ANALYSE DES BESOINS ET MODÈLE DE MENACE

L'apprentissage fédéré a pour objectif de préserver la confidentialité des données en évitant leur centralisation. Toutefois, de nombreux travaux ont démontré que cette propriété n'est pas suffisante en l'absence de mécanismes de sécurité supplémentaires, en raison des fuites possibles à partir des gradients échangés et des communications entre les entités participantes. Cette section illustre de manière précise les besoins de sécurité du système proposé ainsi que le modèle de menace correspondant Yang *et al.* (2022); Redmon *et al.* (2015).

2.1.1 OBJECTIFS DE SÉCURITÉ ET CONFIDENTIALITÉ

Les objectifs de sécurité du mécanisme proposé correspondent aux vulnérabilités identifiées dans la littérature sur l'apprentissage fédéré et les systèmes distribués sécurisés. Le mécanisme doit empêcher toute reconstruction ou inférence d'informations sensibles à partir des gradients échangés durant l'entraînement en particulier, le serveur ne doit jamais avoir accès aux gradients individuels en clair, afin de contrer les attaques de reconstruction de type Deep Leakage from Gradients et les attaques d'inférence statistique Redmon *et al.* (2015); Ultralytics (2024). Les mises à jour envoyées par les clients doivent être protégées contre toute modification malveillante lors de leur transmission, et le serveur doit être en mesure de détecter toute tentative d'altération des gradients notamment dans des scénarios d'empoisonnement de modèle Shor (1999). Il doit également garantir que seules des entités légitimes participent à l'apprentissage fédéré. chaque client et le serveur doivent être authentifiés de manière cryptographiquement robuste, afin de prévenir l'injection de clés ou de mises à jour malveillantes Chen *et al.* (2016, 2017). Enfin, dans un contexte à long terme, les mécanismes cryptographiques utilisés doivent rester sûrs face à des adversaires disposant de capacités quantiques. Ce qui exclut l'utilisation exclusive de schémas classiques tels que RSA ou ECC, vulnérables à l'algorithme de Shor, et motive l'adoption de primitives post-quantiques normalisées ou en cours de normalisation Zhang *et al.* (2021); Beullens *et al.* (2021).

2.1.2 HYPOTHÈSES DU SYSTÈME

Le système étudié repose sur plusieurs hypothèses largement adoptées dans la littérature sur l'apprentissage fédéré sécurisé. Les clients sont honnêtes lorsqu'ils utilisent le mécanisme, mais pourraient être attaqués par des attaques externes qui cherchent à extraire ou à modifier des informations. Le serveur est réputé pour son honnêteté et sa curiosité. Il suit l'algorithme d'agrégation FedAvg, mais peut tenter d'exploiter les gradients reçus pour inférer des données

privées Bonawitz *et al.* (2017); Yang *et al.* (2022). Il est possible que des attaquants externes interceptent les canaux de communication, c'est pourquoi il est nécessaire d'utiliser un chiffrement élevé des communications. Pour des environnements distribués avec des modèles de vision profonds, il est crucial que les méthodes de sécurité respectent les contraintes réalistes de calcul et de bande passante. Ces hypothèses permettent de se concentrer sur les menaces les plus critiques tout en tenant compte des scénarios de déploiement réalistes Tasnim & Qi (2023); Redmon *et al.* (2015).

2.1.3 MODÈLE D'ADVERSAIRE

Le modèle de menace considéré dans ce travail inclut deux catégories principales d'adversaires, en conformité avec les travaux existants sur la sécurité de l'apprentissage fédéré. Adversaire serveur honnête, mais-curieux, le serveur central qui respecte l'algorithme d'agrégation (FedAvg), mais il cherche à exploiter les gradients ou les mises à jour reçues afin d'extraire des informations sensibles sur les données locales des clients. Ce modèle d'adversaire est considéré comme l'un des plus réalistes dans les systèmes FL déployé à grande échelle Yang *et al.* (2018, 2022).

Attaquant externe

Un adversaire externe peut intercepter les communications client-serveur, tenter des attaques de l'homme du milieu, injecter des mises à jour malveillantes ou procéder à des attaques d'empoisonnement ou d'inférence d'appartenance Ultralytics (2024); Dendorfer *et al.* (2020). Cette approche a été proposée afin de pallier les limites des solutions existantes, qui traitent généralement de manière isolée la confidentialité des gradients, l'authentification des participants ou la sécurité des communications, sans offrir une protection globale face aux menaces post-quantiques et aux attaques par exploitation des gradients. En combinant

plusieurs mécanismes complémentaires, l'objectif est d'assurer une protection cohérente et robuste du processus d'apprentissage fédéré. À savoir :

1. Un échange des clés post-quantiques pour la sécurité de la communication.
2. Un chiffrement symétrique des gradients et des modèles.
3. Des signatures numériques pour l'authentification et l'intégrité.
4. Un masquage pairwise qui empêche l'observation des gradients individuels, même par le serveur.

2.2 CHOIX CRYPTOGRAPHIQUES ET JUSTIFICATIONS

La sécurisation de l'apprentissage fédéré proposé repose sur une sélection rigoureuse de primitives cryptographiques adaptées à un environnement distribué, soumis à des contraintes de confidentialité, d'intégrité et de résistance aux menaces post-quantiques. Les choix effectués tiennent compte à la fois de la robustesse cryptographique, de l'efficacité computationnelle et de la compatibilité avec l'entraînement fédéré de modèles de vision profonds. Cette section détaille et justifie les primitives retenues pour l'échange de clés, l'authentification, la dérivation de clés et le chiffrement symétrique Li *et al.* (2024).

2.2.1 CHOIX DE KYBER512 POUR L'ÉCHANGE DE CLÉS

L'échange de clés constitue la première pierre de la sécurité du mécanisme de sécurité, car il conditionne la confidentialité de toutes les communications entre les clients et le serveur. Les schémas cryptographiques classiques fondés sur le logarithme discret ou la factorisation tels que RSA ou ECC sont reconnus comme vulnérables face à un adversaire quantique capable d'exécuter l'algorithme de Shor. Ainsi, leur utilisation n'est donc pas adaptée à des systèmes déployés à long terme. Dans ce contexte, le mécanisme de sécurité utilisé

est CRYSTALS- Kyber qui est un mécanisme d'encapsulation de clés basé sur les réseaux euclidiens et sélectionnés par le NIST dans le cadre du processus de standardisation de la cryptographie post-quantique. Le paramètre Kyber512 offre un niveau de sécurité équivalent à 128 bits tout en limitant la taille des clés et le coût computationnel, ce qui en fait un compromis pertinent pour des environnements distribués impliquant des échanges fréquents Zhang *et al.* (2021); Beullens *et al.* (2021). Dans un contexte d'apprentissage fédéré, le mécanisme d'échange de clés joue un rôle central, car il est exécuté de manière répétée entre des entités hétérogènes et conditionne la sécurité de l'ensemble des communications. Le choix de ce mécanisme doit donc concilier la sécurité post-quantique, l'efficacité computationnelle et la facilité d'intégration dans des protocoles existants. Afin de justifier le choix de Kyber512 dans l'approche proposée, ce mécanisme est évalué au regard de ces contraintes spécifiques. À ce titre :

1. Kyber512 présente plusieurs avantages dans un cadre d'apprentissage fédéré.
2. Des opérations d'encapsulation et de décapsulation rapides.
3. Une intégration simple dans des protocoles existants.
4. Une sécurité fondée sur le problème Learning With Errors, reconnu comme résistant aux attaques quantiques et classiques connues.

Dans l'approche proposée, Kyber512 est utilisé pour établir des clés secrètes partagées entre le serveur et chaque client participant, ainsi qu'entre clients voisins dans le cadre du masquage pairwise. Ces clés constituent une base cryptographique solide pour la sécurisation des gradients échangés et du modèle global assurant ainsi la confidentialité du processus d'apprentissage fédéré Redmon *et al.* (2015); AbhishekV *et al.* (2022).

2.2.2 CHOIX DE CRYSTALS-DILITHIUM POUR LES SIGNATURES

La sécurité d'un système d'apprentissage fédéré ne repose pas uniquement sur la confidentialité des communications. Il est tout aussi essentiel de garantir l'authenticité des différentes entités participant et l'intégrité des différentes mises à jour échangées. Pour répondre à ces exigences dans un contexte post-quantique, le mécanisme repose sur des signatures numériques basées sur CRYSTALS-Dilithium, un schéma de signature qui est à base de réseaux sélectionnés par le NIST dans le cadre du processus de standardisation de la cryptographie post-quantique. Dilithium offre différents niveaux de sécurité, dont une variante alignée sur le niveau 128 bits avec des performances compétitives par rapport aux schémas classiques, tout en assurant une résistance aux attaques quantiques connues Zhang *et al.* (2021); Beullens *et al.* (2021); AbhishekV *et al.* (2022).

Plusieurs avantages sont présentés par CRYSTALS-Dilithium pour un système distribué d'apprentissage fédéré, et il repose sur des problèmes difficiles en théorie des réseaux (Module-LWE/Module-SIS). De plus il dispose d'une implémentation efficace et adaptée à des environnements contraints et fournis des signatures qui sont déterminées en évitant certaines vulnérabilités liées à la génération de nonces, tout en restant compatible avec des échanges fréquents de messages entre clients et serveur Tasnim & Qi (2023).

Dans le mécanisme de sécurité proposé, CRYSTALS-Dilithium est utilisé pour :

1. Signer les clés publiques Kyber lors de l'enrôlement des clients.
2. Authentifier les gradients transmis au serveur.
3. Signer le modèle global après agrégation, avant sa redistribution aux clients.

Ce mécanisme améliore la prévention des attaques de type homme du milieu, des mises à jour malveillantes et de certaines formes d'empoisonnement de modèle, qui ont été identifiées dans la littérature sur l'apprentissage fédéré sécurisé Paillier (1999); Yang *et al.* (2022).

2.2.3 DÉRIVATION DE CLÉS AVEC HKDF

Bien que Kyber512 permette d'établir un secret partagé entre les entités, l'utilisation directe de ces secrets comme clés de chiffrement symétrique n'est pas recommandée. Afin de garantir une séparation stricte des usages cryptographiques et d'améliorer la robustesse globale du mécanisme, une fonction de dérivation de clés est introduite. Il utilise HKDF (HMAC-based Key Derivation Function), une fonction standardisée permettant d'extraire et d'étendre un secret partagé afin de produire des clés cryptographiquement indépendantes. Dans ce contexte, nous avons utilisé HKDF pour dériver les clés symétriques à partir des secrets établis par Kyber512. Cette approche garantit que les clés AES utilisées pour le chiffrement des gradients et du modèle global du serveur sont indépendantes du mécanisme d'échange de clés post-quantique, réduisant ainsi les risques de corrélation et renforçant la sécurité globale du système Gurung *et al.* (2023).

2.2.4 CHIFFREMENT SYMÉTRIQUE AVEC AES-CBC

Une fois que les clés symétriques ont été dérivées, le chiffrement des gradients et du modèle global est assuré à l'aide de l'algorithme AES en mode CBC (Cipher Block Chaining). AES un standard éprouvé est largement utilisé dans les systèmes sécurisés en raison de son efficacité et de sa robustesse lorsqu'il est correctement implémenté. Le mode CBC permet de faire le chiffrement de données de taille variable et garantit que deux messages identiques produisent des ciphertexts différents grâce à l'utilisation d'une (IV) un vecteur d'initialisation aléatoire. Dans le mécanisme proposé, AES-CBC est utilisé pour :

1. Chiffrer les gradients protégés avant leur transmission au serveur.
2. Chiffrer le modèle global avant de le redistribuer aux clients.

Combiné aux signatures post-quantiques CRYSTALS-Dilithium, le chiffrement AES-CBC permet d'assurer simultanément la confidentialité, l'intégrité et l'authenticité des données échangées, tout en maintenant un coût computationnel compatible avec l'entraînement fédéré de modèles de vision profonde Yang *et al.* (2022); AbhishekV *et al.* (2022).

2.3 ARCHITECTURE DE L'APPRENTISSAGE FÉDÉRÉ SÉCURISÉ

Cette section décrit l'architecture globale du système d'apprentissage fédéré sécurisé que nous avons proposé. Elle présente l'organisation entre le client et serveur, la construction du graphe de voisins utilisé pour le masquage pairwise des gradients, ainsi que le flux complet des communications sécurisées. L'objectif est de démontrer comment les primitives cryptographiques et protocolaires mentionnées plus tôt s'intègrent de manière cohérente dans une architecture opérationnelle et réaliste Bonawitz *et al.* (2017).

2.3.1 ARCHITECTURE GLOBALE CLIENT–SERVEUR

Notre système repose sur une architecture client-serveur classique figure 2.1, communément adoptée dans les déploiements d'apprentissages fédérés à grande échelle. Dans cette architecture, le serveur central assume un rôle d'orchestrateur, tandis que les clients conservent leurs données localement et effectuent les phases d'entraînement. Dans ce système le rôle principal du serveur est :

1. D'initialiser le modèle global de détection d'objets pour un début.
2. De distribuer ce modèle aux clients au début de chaque round fédéré.
3. De vérifier la signature du modèle reçu des clients.
4. De faire une agrégation de ces mises à jour via l'algorithme FedAvg.
5. De signer et de chiffrer ce modèle agrégé.

6. Redistribuer le modèle global mis à jour.

Les clients quant à eux sont responsables des opérations locales effectuées après la réception du modèle initial transmis par le serveur, notamment :

1. De vérifier la signature du modèle.
2. De déchiffrer le modèle, puis mets à jour son modèle.
3. L'entraînement local du modèle sur leurs données privées.
4. Appliquer le masquage pairwise et bruit annulable lors de l'agrégation.
5. Signer et chiffrer le modèle avant d'envoyer le modèle au serveur.

Cette séparation stricte des rôles permet d'assurer que les données brutes ne quittent jamais les dispositifs des clients, tout en autorisant une mise à jour collaborative du modèle global, conformément aux principes fondamentaux de l'apprentissage fédéré Zuo *et al.* (2022); Yang *et al.* (2022); AbhishekV *et al.* (2022).

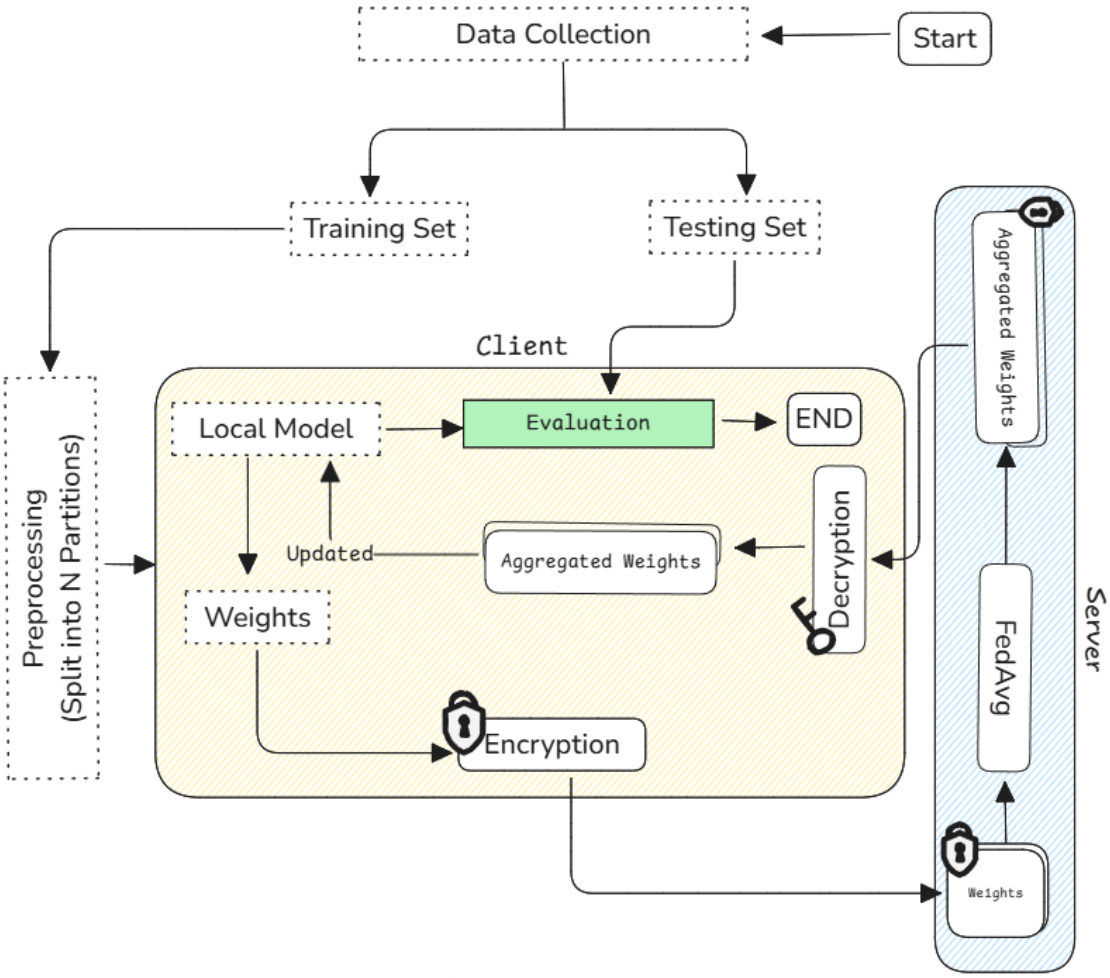

FIGURE 2.1 : Architecture generale

2.3.2 GRAPHE DE VOISINS POUR LE MASQUAGE PAIRWISE

Afin de protéger les gradients individuels contre toute tentative d'observation par un serveur honnête, mais curieux, cette approche adopte un mécanisme de masquage pairwise basé sur un graphe de voisins, inspiré des approches générales de confidentialité et d'agrégation sécurisée en FL Zhang *et al.* (2021). Après l'enregistrement et la vérification des clés publiques des clients, le serveur génère un graphe aléatoire reliant chaque client à un nombre restreint de voisins. Ainsi, ce nombre de voisins est choisi pour :

1. Réduire la complexité des échanges,
2. Limiter le surcoût computationnel,
3. Garantir l'annulation globale des masques lors de l'agrégation.

Pour chaque paire de clients voisins, Kyber512 établit une clé secrète partagée Zuo *et al.* (2022); Yang *et al.* (2022). Ces clés secrètes sont ensuite utilisées pour générer localement des masques symétriques de signe opposé qui sont appliqués aux gradients avant l'envoi. Par conception mathématique, la somme des masques de tous les clients est égale à zéro, ce qui garantit qu'au moment de l'agrégation, les masques s'annulent et que le serveur obtient uniquement la somme globale des gradients, sans jamais pouvoir isoler une contribution individuelle. Ce mécanisme protège efficacement contre les attaques de reconstruction de gradients, en particulier celles analysées dans les travaux sur la fuite profonde des gradients DLG (Deep Leakage from Gradients) Zhu *et al.* (2019), ainsi que dans les analyses générales des vulnérabilités et contre-mesures en FL Zhang *et al.* (2021).

2.3.3 FLUX DES COMMUNICATIONS SÉCURISÉES

À chaque ronde fédérée, les communications sécurisées entre le serveur et les clients suivent une séquence d'opérations bien définie, intégrant de manière systématique les différents mécanismes cryptographiques présentés dans les sections précédentes, afin de garantir la confidentialité, l'intégrité et l'authenticité des échanges Bonawitz *et al.* (2017) telles que :

1. La distribution du modèle global : Le serveur encapsule d'abord une clé sécrète partagée avec chaque client à l'aide de CRYSTALS Kyber512, puis dérive cette clé via HKDF SHA256, une clé de chiffrement symétrique et une clé HMAC distincte. Le modèle global est chiffré en utilisant AES CBC avec un vecteur d'initialisation (IV) aléatoire de 16 octets généré pour chaque message, puis un HMAC SHA256 est calculé sur le

ciphertext selon un schéma (Encrypt then MAC). Le paquet résultant (IV, ciphertext, tag HMAC) est finalement signé avec la clé privée du CRYSTALS Dilithium du serveur avant d'être transmis à chaque client Chen *et al.* (2016).

2. La vérification et entraînement local : Chaque client vérifie d'abord la signature via CRYSTALS Dilithium pour s'assurer que le modèle reçu est authentique, puis vérifie le tag HMAC avant de déchiffrer les poids à l'aide de la clé AES CBC qui est dérivée de la clé sécrète Kyber512 via HKDF. Une fois le modèle reconstruit en clair, le client effectue son entraînement local sur ses données privées Joseph *et al.* (2022).
3. La protection des gradients : À l'issue de cet entraînement, chaque client applique d'abord un masquage pairwise à ses différents gradients au sein d'un graphe de voisins, puis ajoute un bruit gaussien déterministe qui est destiné à renforcer la résistance aux attaques d'inférence et aux attaques de reconstruction. Ces deux mécanismes garantissent que les gradients individuels ne peuvent pas être isolés par un serveur honnête mais curieux Bonawitz *et al.* (2017); Yang *et al.* (2022).
4. Chiffrement et transmission : Les gradients protégés sont chiffrés à l'aide de AES CBC avec une clé de chiffrement dérivée via Kyber512 et HKDF. Pour chaque message a envoyer, on génère un vecteur d'initialisation (IV) aléatoire de 16 octets, puis un HMAC SHA256 est calculé sur le ciphertext dans un schéma Encrypt then MAC avec une clé MAC dédiée. L'IV, le ciphertext et le tag HMAC sont combinés pour former le message final, qui est ensuite signé avec CRYSTALS Dilithium avant transmission au serveur Joseph *et al.* (2022).
5. Vérification et agrégation : Le serveur vérifie l'authenticité et l'intégrité des mises à jour reçues en utilisant la signature CRYSTALS Dilithium et tag HMAC, puis déchiffre les gradients protégés à l'aide des clés dérivées via Kyber512 et HKDF. Ensuite fait l'agrégation des contributions par FedAvg ; grâce à l'annulation mathématique des

masques pairwise, seule la somme agrégée est disponible, sans exposition de gradients individuels Bonawitz *et al.* (2017); Yang *et al.* (2022).

C'est la garantit que les gradients et le modèle global sont protégés à chaque étape du mécanisme de sécurité, tant contre les attaques externes sur le canal de communication que contre les tentatives d'inférence internes au serveur, tout en maintenant un comportement d'apprentissage comparable à un FL non sécurisé Zhang *et al.* (2021); AbhishekV *et al.* (2022). Cette architecture assure une intégration cohérente entre apprentissage fédéré, cryptographie post quantique et mécanismes de protection statistique des gradients, tout en restant compatible avec les contraintes pratiques des modèles de détection d'objets YOLOv8 et SSD utilisés dans le cadre expérimental Sawatzky *et al.* (2019).

2.4 PROTECTION DES GRADIENTS

Dans l'apprentissage fédéré, la protection des gradients locaux est un enjeu fondamental, car plusieurs travaux ont démontré que les gradients transmis au serveur peuvent être exploités pour reconstruire les données d'entraînement ou inférer des informations sensibles, en particulier dans le cas de modèles de vision par ordinateur à forte capacité comme YOLO Paillier (1999); Redmon *et al.* (2015). Afin de répondre à ces menaces, l'approche proposée combine le masquage pairwise, l'annulation mathématique des masques et l'ajout de bruit gaussien, assurant une protection renforcée contre les attaques tout en conservant la convergence du modèle Bonawitz *et al.* (2017); Beullens *et al.* (2021).

2.4.1 MASQUAGE PAIRWISE DES GRADIENTS

Le masquage pairwise repose sur une propriété d'anti-symétrie garantissant que, pour toute paire de clients (i, j) participant à un même round fédéré, les masques générés sont opposés deux à deux. Cette propriété est formalisée par l'équation suivante :

$$m_{i,j}^{(t)} = -m_{j,i}^{(t)}, \tag{2.1}$$

où $m_{i,j}^{(t)}$ désigne le masque appliqué par le client i pour le voisin j au round t. Cette relation est assurée par l'utilisation d'une clé symétrique partagée identique $K_{i,j} = K_{j,i}$ et par l'affectation d'un signe opposé selon l'ordre des indices.

À chaque round fédéré t, chaque client entraîne localement un modèle et calcule un gradient brut noté $g_i^{(t)} \in \mathbb{R}^d$, où d correspond au nombre total de paramètres du modèle. Le serveur définit une topologie de communication sous la forme d'un graphe de voisins et attribue à chaque client i un ensemble restreint de voisins $\mathcal{N}(i)$.

Pour chaque paire de clients (i, j) appartenant au graphe de voisins, une clé symétrique partagée $K_{i,j}$ est établie à l'aide d'un mécanisme d'encapsulation de clés post-quantique. À partir de cette clé, un masque pairwise est dérivé à l'aide d'une fonction cryptographique déterministe telle qu'une fonction HMAC, dépendant du numéro du round :

$$m_{i,j}^{(t)} = f(K_{i,j}, \text{nonce}^{(t)}), \tag{2.2}$$

où $\text{nonce}^{(t)}$ est une valeur dépendante du round, garantissant l'unicité des masques entre itérations successives.

Le masque total appliqué par le client i est alors obtenu en sommant l'ensemble des masques associés à ses voisins :

$$M_i^{(t)} = \sum_{j \in \mathcal{N}(i)} m_{i,j}^{(t)}. \tag{2.3}$$

Le gradient transmis au serveur est finalement obtenu en combinant le gradient brut et le masque total :

$$\widetilde{g}_i^{(t)} = g_i^{(t)} + M_i^{(t)}. \tag{2.4}$$

Grâce à la propriété d'anti-symétrie définie à l'équation 2.1, la somme des masques appliqués par l'ensemble des clients s'annule lors de l'agrégation globale, empêchant le serveur d'accéder aux gradients individuels en clair Bonawitz *et al.* (2017).

```
[LOG] Masque couche 177 avec voisin 3 (signe 1)
[LOG] Masque couche 177 avec voisin 4 (signe 1)
[LOG] Masque couche 178 avec voisin 1 (signe -1)
[LOG] Masque couche 178 avec voisin 3 (signe 1)
[LOG] Masque couche 178 avec voisin 4 (signe 1)
[LOG] Masque couche 179 avec voisin 1 (signe -1)
[LOG] Masque couche 179 avec voisin 3 (signe 1)
[LOG] Masque couche 179 avec voisin 4 (signe 1)
[LOG] Masque couche 180 avec voisin 1 (signe -1)
[LOG] Masque couche 180 avec voisin 3 (signe 1)
[LOG] Masque couche 180 avec voisin 4 (signe 1)
[LOG] Masque couche 181 avec voisin 1 (signe -1)
[LOG] Masque couche 181 avec voisin 3 (signe 1)
[LOG] Masque couche 181 avec voisin 4 (signe 1)
[LOG] Masque couche 182 avec voisin 1 (signe -1)
```

FIGURE 2.2 : Génération des masques pairwise couche par couche et par voisin côté client

Cette figure 2.2 met en évidence le mécanisme de génération des masques pairwise côté client. Pour chaque couche du modèle et pour chacun des voisins définis dans le graphe, un masque est généré localement avec un signe positif ou négatif. Cette affectation de signes opposés garantit l'annulation mathématique des masques lors de l'agrégation sécurisée, tout en empêchant le serveur d'accéder aux gradients individuels en clair.

2.4.2 ANNULATION MATHÉMATIQUE DES MASQUES

Ainsi, lors de l'agrégation globale, la somme des masques appliqués par l'ensemble des clients satisfait :

$$\sum_{i=1}^{K} \sum_{j \in \mathcal{N}(i)} m_{i,j}^{(t)} = 0 \tag{2.5}$$

L'équation 2.5 illustre que les contributions des masques s'annulent deux à deux, de sorte que le serveur ne peut accéder qu'au gradient agrégé sans révéler les gradients individuels des clients. Cette propriété nous garantit que lors de l'agrégation globale des modèles, les masques s'annulent exactement et n'affectent pas la mise à jour du modèle. Les logs expérimentaux montrent explicitement cette symétrie via les signes opposés appliqués aux masques couche par couche pour les voisins correspondants, confirmant la correction de l'implémentation. Cette propriété nous garantit que lors de l'agrégation globale des modèles, les masques s'annulent exactement et n'affectent pas la mise à jour du modèle, conformément aux mécanismes de sécure agrégation basée sur le masquage pairwise et aux travaux sur la protection des gradients en apprentissage fédéré. Les logs expérimentaux montrent explicitement cette symétrie via les signes opposés appliqués aux masques couche par couche pour les voisins correspondants, confirmant la correction de l'implémentation Bonawitz *et al.* (2017); Redmon *et al.* (2015).

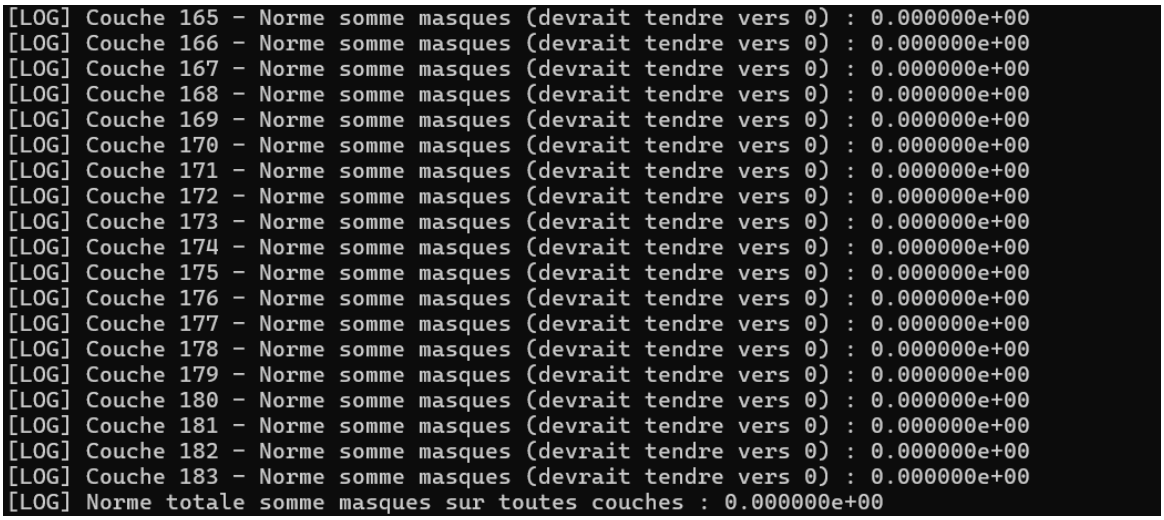

FIGURE 2.3 : Vérification couche par couche de l'annulation des masques pairwise lors de l'agrégation sécurisée

Cette figure 2.3 illustre la norme de la somme des masques appliqués aux gradients pour chaque couche du modèle. La norme nulle observée confirme l'annulation mathématique des masques et la validité du mécanisme d'agrégation FedAvg sécurisée.

2.4.3 AJOUT DE BRUIT GAUSSIEN

Pour renforcer la protection contre les attaques statistiques et les attaques d'inférence inter-rounds, chaque client ajoute un bruit gaussien indépendant à son gradient masqué. Le bruit gaussien ajouté localement par chaque client est modélisé par.

$$\varepsilon_i^{(t)} \sim \mathcal{N}(0, \sigma^2 I) \tag{2.6}$$

L'Equation 2.6 indique que le bruit est centré d'espérance nulle et de variance σ^2, supposé indépendant et isotrope, ce qui permet de perturber les gradients sans introduire de biais systématique.

Le gradient transmis au serveur devient alors.

$$g_i'^{(t)} = \widetilde{g}_i^{(t)} + \varepsilon_i^{(t)} \tag{2.7}$$

L'équation 2.7 formalise l'ajout du bruit au gradient déjà masqué, renforçant ainsi la protection contre les attaques de reconstruction. L'impact de ce mécanisme est évalué expérimentalement à l'aide de métriques de similarité statistique. Pour le gradient masqué et bruité, on observe :

$$\mathrm{sim}_{\cos}\left(g_i^{(t)}, g_i'^{(t)}\right) \approx 0{,}058, \quad \text{Spearman} \approx 0{,}039, \quad \text{Kendall} \approx 0{,}026 \tag{2.8}$$

Ces métriques mettent en évidence une corrélation extrêmement faible entre le gradient original et le gradient protégé, confirmant l'efficacité conjointe du masquage et de l'ajout de bruit pour réduire significativement la fuite d'information.

Par ailleurs, l'entropie du gradient protégé atteint environ $H\left(g_i'^{(t)}\right) \approx 6{,}12$, contre $H\left(g_i^{(t)}\right) \approx 1{,}59$ pour le gradient en clair. Ces valeurs traduisent une rupture quasi totale de la structure informationnelle entre les gradients protégés et les gradients d'origine, en cohérence avec les travaux existants sur les fuites de gradients et l'effet de l'ajout de bruit pour renforcer la confidentialité en apprentissage fédéré Redmon *et al.* (2015); Frankel *et al.* (2003).

2.4.4 IMPACT SUR LA CONFIDENTIALITÉ ET LA CONVERGENCE

Le serveur agrège les gradients protégés en utilisant l'algorithme FedAvg McMahan *et al.* (2017).

$$g_{\text{agg}}^{(t)} = \sum_{i=1}^{K} \frac{n_i}{n} g_i'^{(t)} \tag{2.9}$$

où $g_i'^{(t)}$ désigne le gradient masqué et bruité du client i au round t, n_i la taille de son jeu de données local, et $n = \sum_{i=1}^{K} n_i$ la taille totale des données comme indiquées dans la figure 2.9. Cette étape d'agrégation permet de combiner les contributions locales tout en limitant l'exposition des gradients individuels, connus pour être vulnérables aux attaques d'inférence Zhu *et al.* (2019).

Cette équation 2.10 signifie que le modèle global est mis à jour en se déplaçant dans la direction opposée au gradient agrégé des clients, selon une étape de descente de gradient de (n) pas McMahan *et al.* (2017).

$$w^{(t+1)} = w^{(t)} - \eta\, g_{\text{agg}}^{(t)} \tag{2.10}$$

L'équation 2.11 montre en développant l'expression du gradient agrégé, on obtient :

$$g_{\text{agg}}^{(t)} = \sum_{i=1}^{K} \frac{n_i}{n} \left(g_i^{(t)} + M_i^{(t)} + \varepsilon_i^{(t)} \right) \tag{2.11}$$

Cette décomposition met en évidence l'ajout de deux mécanismes de protection, figure 2.12. Les masques pairwise $M_i^{(t)}$ destinés à empêcher l'observation des gradients locaux et le bruit gaussien $\varepsilon_i^{(t)}$ visant à réduire les risques de reconstruction Zhu *et al.* (2019).

Cette expression peut être réécrite sous la forme suivante :

$$g_{\text{agg}}^{(t)} = \underbrace{\sum_{i=1}^{K} \frac{n_i}{n} g_i^{(t)}}_{\text{gradient FedAvg classique}} \quad \underbrace{\sum_{i=1}^{K} \frac{n_i}{n} M_i^{(t)}}_{\approx 0} \underbrace{\sum_{i=1}^{K} \frac{n_i}{n} \varepsilon_i^{(t)}}_{\text{bruit moyen}} \tag{2.12}$$

L'équation 2.13 montre que les masques pairwise s'annulent globalement et que le bruit gaussien est nul, garantissant que seule l'information utile pour l'apprentissage est conservée lors de l'agrégation.

$$\sum_{i=1}^{K} \frac{n_i}{n} M_i^{(t)} = 0 \quad \text{et} \quad \mathbb{E}\left[\sum_{i=1}^{K} \frac{n_i}{n} \varepsilon_i^{(t)}\right] = 0 \tag{2.13}$$

L'algorithme retrouve en espérance une dynamique de mise à jour proche de FedAvg classique, ce qui est cohérent avec les analyses de convergence de FedAvg sur données non-IID. Les résultats expérimentaux montrent que, malgré la forte perturbation introduite sur les gradients :

1. Le modèle converge normalement ;
2. Les pertes de détection (box loss, cls loss, dfl loss) diminuent de manière stable ;
3. Les performances restent élevées.

Résultats YOLOv8n observés (rounds fédérés sécurisés) :

1. mAP50 $\approx 0{,}90$;
2. mAP50–95 $\approx 0{,}75$;
3. précision $P \approx 0{,}88$ et
4. rappel $R \approx 0{,}80$.

Ces résultats confirment que la protection utilisée n'empêche pas l'apprentissage efficace du modèle, même dans un contexte de données (non-IID) et de modèle profond de vision par ordinateur, en ligne avec les travaux sur la convergence de FedAvg et les défenses contre

les attaques de reconstruction par bruitage des gradients Kaur & Singh (2022); Frankel *et al.* (2003).

2.5 SÉCURISATION DES COMMUNICATIONS ET DES MODÈLES

Dans un système d'apprentissage fédéré sécurisé, la protection des gradients doit obligatoirement être complétée par la sécurisation rigoureuse des communications et du modèle global. En effet, même en présence de mécanismes d'agrégation sécurisée, des attaques visant l'interception, la modification ou la substitution des messages peuvent compromettre la confidentialité, l'intégrité ainsi que la convergence du modèle global. L'approche proposée repose sur une protection post-quantique des gradients combinant masquage pairwise, chiffrement hybride, signatures numériques et vérifications cryptographiques systématiques, afin de garantir la sécurité de bout en bout des échanges client-serveur Shor (1999); Yang *et al.* (2022).

2.5.1 CHIFFREMENT ET SIGNATURE DES GRADIENTS

Après l'étape de la protection locale des gradients (masquage pairwise et ajout de bruit gaussien), chaque client obtient un gradient protégé noté $g_i'^{(t)}$. Avant la transmission au serveur, ce gradient est sécurisé au niveau cryptographique.

Une clé symétrique k_i, dérivée à partir du secret partagé issu de l'échange Kyber512 via HKDF, est utilisée pour chiffrer le gradient protégé à l'aide d'AES-CBC comme indiqué dans l'équation 2.14.

$$c_i^{(t)} = \text{AES-CBC\_Enc}\left(k_i, g_i'^{(t)}, \text{IV}_i^{(t)}\right) \tag{2.14}$$

où ($IV_i^{(t)}$) représente un vecteur d'initialisation aléatoire unique pour chaque message. Ce chiffrement garantit la confidentialité des gradients, même en cas d'interception du canal de communication AbhishekV *et al.* (2022).

Afin d'assurer l'authenticité et l'intégrité des mises à jour locales, chaque ciphertext ($c_i^{(t)}$) est ensuite signé à l'aide de la clé privée Dilithium du client, conformément à l'équation 2.15

$$\sigma_i^{(t)} = \text{Signature}_{\text{Dilithium}}\left(sk_i, c_i^{(t)}\right) \tag{2.15}$$

où sk_i désigne la clé secrète de signature du client i. Cette signature permet de détecter toute modification non autorisée et protège le système contre les attaques de type Man-in-the-Middle ou l'injection de gradients malveillants Shor (1999); Chen *et al.* (2016).

2.5.2 VÉRIFICATION CÔTÉ SERVEUR

À la réception des modèles des clients, le serveur applique une procédure de vérification stricte avant toute opération d'agrégation. Tout d'abord, la signature associée à chaque message est vérifiée à l'aide de la clé publique Dilithium du client conformément à l'équation 2.16.

$$\text{Verify}_{\text{Dilithium}}\left(pk_i, c_i^{(t)}, \sigma_i^{(t)}\right) \tag{2.16}$$

Où pk_i désigne la clé publique de signature du client i, $c_i^{(t)}$ le ciphertext reçu et $\sigma_i^{(t)}$ la signature associée. Tout message dont la signature est invalide est rejeté, ce qui permet d'exclure les contributions falsifiées ou provenant d'entités non légitime Shor (1999); Chen *et al.* (2016). Une fois que cette signature est validée, le serveur procède au déchiffrement du

gradient protégé à l'aide de la clé symétrique correspondante comme indiqué dans l'équation 2.17.

$$g_i'^{(t)} = \text{Dechiffrement\_AES-CBC}\left(k_i, c_i^{(t)}, \text{IV}_i^{(t)}\right) \tag{2.17}$$

Le serveur n'accède jamais aux gradients bruts $g_i^{(t)}$, mais uniquement aux gradients masqués et bruités $g_i'^{(t)}$, ce qui est conforme au modèle de menace du serveur honnête, mais-curieux, largement adopté dans la littérature sur l'apprentissage fédéré sécurisé Paillier (1999). Symétriquement à chaque round, les clients vérifient la signature Dilithium apposée sur le modèle global agrégé avant de le déchiffrer et de l'utiliser pour la phase d'entraînement local, ce qui empêche l'injection de modèles globaux altérés et renforce la chaîne de confiance bout en bout Shor (1999); Chen *et al.* (2016).

2.5.3 SIGNATURE, CHIFFREMENT ET REDISTRIBUTION DU MODÈLE GLOBAL

Après la phase d'agrégation FedAvg, le serveur obtient un modèle global mis à jour. Avant de le redistribuer aux clients ce modèle est également sécurisé. Le serveur commence par signer le modèle global $w^{(t+1)}$ à l'aide de sa clé privée Dilithium comme indiqué dans l'équation 2.18.

$$\sigma_{\text{srv}}^{(t+1)} = \text{Signature}_{\text{Dilithium}}\left(sk_{\text{srv}}, w^{(t+1)}\right) \tag{2.18}$$

Où sk_{srv} désigne la clé secrète de signature du serveur. Cette signature garantit à chaque client que le modèle reçu provient bien du serveur légitime et qu'il n'a subi aucune altération Shor (1999). Ensuite, Comme indiqué dans l'équation 2.19 pour chaque client i, le modèle

global est chiffré en utilisant la clé symétrique dédiée k_i, dérivée à partir de la clé secrète post-quantique établi avec ce client (Kyber512 + HKDF).

$$c_{\text{srv}\to i}^{(t+1)} = \text{AES-CBC\_Enc}\left(k_i, w^{(t+1)}, \text{IV}_{\text{srv}\to i}^{(t+1)}\right) \tag{2.19}$$

où $\text{IV}_{\text{srv}\to i}^{(t+1)}$ est un vecteur d'initialisation (IV) aléatoire unique pour ce chiffrement, ce qui évite que deux modèles identiques produisent les mêmes ciphertexts est il renforce la confidentialité sur le canal AbhishekV *et al.* (2022). À la réception, chaque client vérifie la signature du serveur, déchiffre le modèle global et charge les nouveaux poids afin d'initier le round fédéré suivant. Ce mécanisme empêche l'interception, la falsification ou l'injection de modèles malveillants et assure la continuité sécurisée du processus d'apprentissage fédéré Yang *et al.* (2022).

2.6 CONCLUSION DU CHAPITRE

Ce chapitre a présenté en détail toute l'approche proposée pour la mise en œuvre d'un système d'apprentissage fédéré sécurisé qui résiste aux menaces post-quantiques, conçu spécifiquement pour des tâches de détection d'objets qui sont basées sur des modèles de vision profonde. L'approche adoptée s'inscrit dans la logique de sécurité de bout en bout, couvrant ainsi à la fois la protection des gradients locaux, la sécurisation des communications et l'intégrité du modèle global. L'analyse des besoins et du modèle de menace a permis d'identifier les vulnérabilités majeures inhérentes à l'apprentissage fédéré notamment la fuite d'informations par les gradients, les attaques actives telles que l'empoisonnement de modèle, ainsi que les risques liés à un serveur honnête, mais-curieux ou à un attaquant externe Paillier (1999); Shor (1999). Ces constats ont incité à faire un choix d'un mécanisme de sécurité combinant plusieurs outils complémentaires plutôt qu'une solution isolée. L'approche est basée sur des

choix cryptographiques justifiés et conformes aux recommandations actuelles. En utilisant de Kyber512 il permet de créer des clés secrètes partagés résistants aux attaques quantiques, tandis que la dérivation de clés via HKDF garantit l'indépendance cryptographique des clés utilisées pour le chiffrement symétrique. Les signatures post-quantiques CRYSTALS-Dilithium permettant une authentification une intégrité parfaite des messages échangés, et le chiffrement AES-CBC protège efficacement la confidentialité des gradients et du modèle global durant leur transit Zhang *et al.* (2021); Beullens *et al.* (2021); Chen *et al.* (2016). La protection des gradients est un élément clé du mécanisme de sécurité proposé. Le masquage pairwise, inspiré des travaux de sécure agrégation de Bonawitz et al, empêche toute observation directe des gradients individuels par le serveur, tandis que l'ajout de bruit gaussien réduit significativement les risques d'attaques de reconstruction et d'inférence statistique.L'annulation mathématique des masques assure que ces mécanismes de protection ne perturbent pas la mise à jour globale du modèle, ce qui préservant ainsi la convergence de l'apprentissage Bonawitz *et al.* (2017); Redmon *et al.* (2015). Enfin, la sécurité complète des communications et la redistribution protégée du modèle global assurent la continuité et la fiabilité du processus fédéré sur plusieurs rounds d'entraînement. En utilisant la démarche méthodologique présentée, ce travail établit un cadre cohérent et opérationnel pour l'apprentissage fédéré sécurisé en contexte post-quantique, tout en restant compatible avec des modèles de vision par ordinateur complexes tels que (YOLOv8, YOLO11n, SSDLite, . . .). L'approche proposée constitue le fondement des chapitres suivants, consacrés à l'implémentation expérimentale, à l'analyse des résultats et à l'évaluation approfondie de l'impact des mécanismes de sécurité proposés sur la confidentialité, la robustesse et les performances du modèle fédéré Tasnim & Qi (2023).

CHAPITRE III

IMPLEMENTATION ET DEVELOPPEMENT

Ce chapitre présente l'implémentation concrète du mécanisme de sécurité post-quantique introduit au chapitre précédent. Les outils logiciels retenus pour l'orchestration de l'apprentissage fédéré et l'entraînement des modèles, ainsi que la configuration spécifique du modèle YOLOv8 et autres modèles pour la détection d'objets dans un contexte distribué et sécurisé.

3.1 ENVIRONNEMENT DE DÉVELOPPEMENT

L'environnement de développement utilisé a été conçu afin de refléter un scénario réaliste d'apprentissage fédéré tout en tenant compte des contraintes de calcul, de communication et de sécurité qui est propre aux systèmes distribués modernes.

3.1.1 OUTILS LOGICIELS UTILISÉS (PYTORCH, FLOWER, ULTRALYTICS)

L'implémentation logicielle est principalement basée sur deux outils complémentaires. PyTorch est employé pour définir, entraîner et mettre à jour la détection d'objets. Cette approche offre une grande flexibilité pour la manipulation des gradients, l'accès aux paramètres du modèle et l'intégration de mécanismes cryptographiques personnalisés Chen *et al.* (2016). Flower est utilisé comme outil d'orchestration de l'apprentissage fédéré. Il simplifie la coordination des rounds fédérés, la communication entre le serveur et les clients, ainsi que l'implémentation de stratégies d'agrégation telles que FedAvg. Flower offre aussi une intégration transparente avec PyTorch, ce qui facilite le déploiement du mécanisme proposé dans un environnement fédéré réel Tasnim & Qi (2023); AbhishekV *et al.* (2022). et la bibliothèque Ultralytics est employée pour l'implémentation et la configuration des modèles

de détection d'objets de type YOLO, notamment YOLOv8. Elle fournit une interface de haut niveau facilitant l'entraînement, l'évaluation et l'inférence des modèles, tout en reposant sur l'infrastructure PyTorch. L'architecture logicielle combine ainsi la modularité de PyTorch avec les capacités de gestion distribuée de Flower, tout en laissant la possibilité d'insérer des étapes de chiffrement, de signature et de vérification sans modifier le cœur des outils.

3.1.2 CONFIGURATION DE YOLOV8

Les modèles YOLOv8 ont été sélectionnés pour les expérimentations en raison de ses performances élevées en détection d'objets en temps réel et de son récent développement dans la littérature et les applications industrielles. Dans ce travail, la version YOLOv8n est privilégiée, car elle offre un compromis adapté entre précisions, rapidité d'inférence et coût computationnel, ce qui est essentiel dans un contexte d'apprentissage fédéré où les ressources clients peuvent être limitées. La configuration du modèle inclut :

1. Une taille d'entrée qui est adaptée a la capacité matérielle des clients.
2. Un nombre réduit d'époques par rond fédéré afin de limiter le temps d'entraînement local.
3. Un batch size ajusté pour prévenir les dépassements de mémoire.
4. Une initialisation des poids qui est toujours issue du modèle global agrégé reçu du serveur.

Les choix d'environnement, d'outils logiciels et de configuration permettent de garantir une mise en œuvre fidèle à l'approche définie, tout en offrant une évaluation réaliste des performances et des surcoûts induits par les mécanismes de sécurité proposés Chen *et al.* (2016); Abdi (2007).

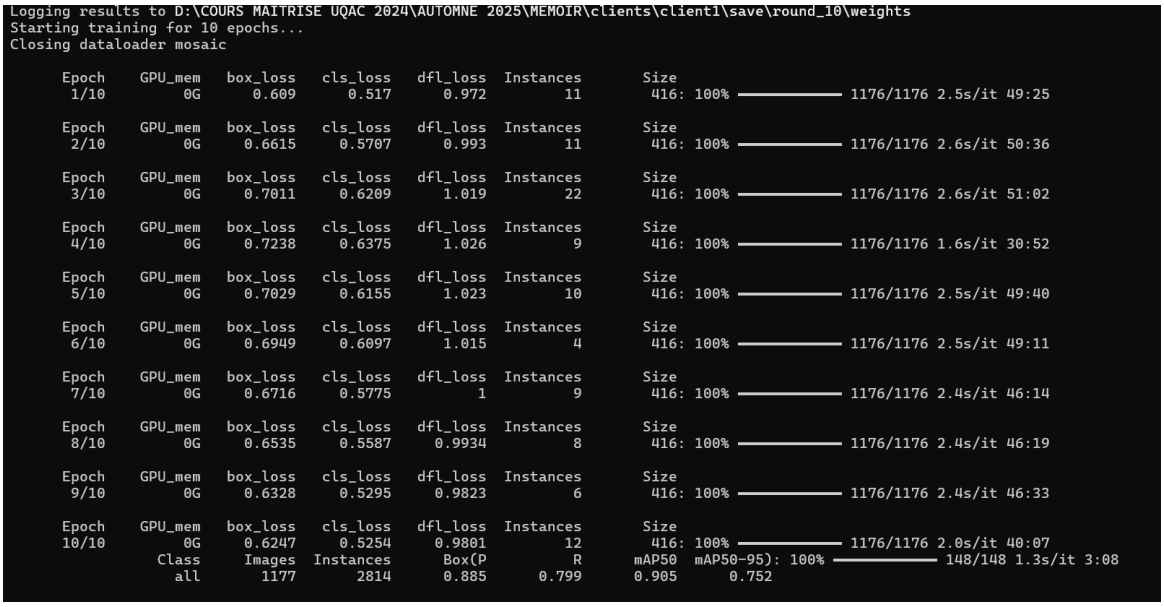

FIGURE 3.1 : Configuration et entrainement du model

3.2 IMPLÉMENTATION CÔTÉ CLIENT

Cette section figure 3.1 décrit l'implémentation détaillée des opérations effectuées sur chaque client participant à l'apprentissage fédéré. Le rôle de chaque client est central, puisqu'il est responsable à la fois de l'entraînement local du modèle de détection d'objets et de l'application des mécanismes cryptographiques qui garantissent la confidentialité et l'intégrité des mises à jour envoyées au serveur.

3.2.1 GÉNÉRATION DES CLÉS CRYPTOGRAPHIQUES

Au démarrage, chaque client génère localement les paires de clés cryptographiques nécessaires aux échanges sécurisés. Deux types de clés sont produits.

1. Une paire de clés post-quantiques Kyber512 (une clé publique et une clé privée) est générée pour l'établissement des clés secrètes partagées résistant aux attaques quantiques.

La clé publique Kyber est destinée à être partagée avec le serveur et les clients voisins définis dans le graphe, tandis que la clé privée reste strictement locale au client et ne sera pas partagé Zhang *et al.* (2021); Beullens *et al.* (2021).

2. Chaque client génère une paire de clés Cristall-Dilitium qui sera utilisée à la signature numérique. Ces clés permettent d'authentifier les messages transmis et de garantir leur intégrité. La clé publique Cristall-Dilitium est envoyée au serveur, accompagnée de la signature de la clé publique Kyber, ce qui permet au serveur de vérifier l'authenticité des clés reçues et d'empêcher toute usurpation d'identité Chen *et al.* (2016).

3.2.2 ENTRAÎNEMENT LOCAL DU MODÈLE YOLO

Nous entraînons localement côté client des modèles de détection d'objets adaptés à un contexte d'apprentissage fédéré contraint en ressources figure 3.1. Pour tester la généralisabilité du pipeline, on étudie trois architectures complémentaires YOLOv8n, YOLOv11n et SSDLite.

Pour les modèles de type (YOLO), la détection est formulée comme une régression en un seul passage, prédisant simultanément les boîtes englobantes et les probabilités de classes, ce qui offre un bon compromis entre vitesse et précision Redmon *et al.* (2015); Zhang *et al.* (2022).

À chaque round FL, la fonction fit() charge d'abord les poids globaux envoyés par le serveur, exécute ensuite l'entraînement local sur la partition privée du client, puis extrait les paramètres du modèle selon notre pipeline (conversion en tenseurs/objets NumPy) afin de faciliter leur transmission et l'agrégation côté serveur McMahan *et al.* (2017); Beutel *et al.* (2020).

Selon le paradigme FL, les données brutes (images et annotations) demeurent strictement locales et ne sont jamais envoyées au serveur ; seules les valeurs mises à jour sont

communiquées. Enfin, nous enregistrons la durée d'entraînement locale, la taille du dataset local et les données de coût pour l'analyse expérimentale. Moore & Corso (2020)

3.2.3 MASQUAGE, BRUITAGE ET CHIFFREMENT DES GRADIENTS

Avant d'envoyer les gradients au serveur, le client applique successivement les mécanismes de protection définis dans la méthodologie.

Tout d'abord, le client calcule les masques pairwise à partir des clés symétriques qu'il a partagées avec ses voisins, dérivées via Kyber512. Ces masques sont générés à l'aide d'une fonction HMAC-SHA256 qui est ajoutée aux gradients locaux, conformément au mécanisme de sécure agrégation Paillier (1999).

Ensuite, un bruit gaussien annulable aléatoire est ajouté aux gradients masqués. Ce bruit permet de réduire la corrélation statistique entre les gradients transmis et les données d'entraînement locales, qui renforce ainsi la résistance aux attaques d'inférence Chen *et al.* (2016).

Les gradients protégés sont ensuite sérialisés, chiffrés à l'aide d'AES-CBC avec une clé dérivée via HKDF, puis signés à l'aide de la clé privée Cristall-Dilitium du client. Cette combinaison permet de garantir la confidentialité des gradients, leur intégrité ainsi que l'authenticité de l'émetteur, même en présence d'un attaquant réseau actif Chen *et al.* (2016).

L'implémentation côté client assure ainsi que les gradients transmis au serveur ne révèlent aucune information exploitable sur les données locales, tout en restant compatibles avec un processus d'agrégation fédérée efficace.

3.3 IMPLÉMENTATION CÔTÉ SERVEUR

Cette section décrit l'implémentation des différentes fonctions assurées par le serveur central dans le cadre de l'apprentissage fédéré sécurisé. Le serveur agit comme coordinateur des rounds fédérés, il est responsable de l'agrégation des mises à jour locales, et garant de la validité cryptographique des différentes contributions reçues. Le modèle de menace considéré suppose un serveur honnête,-mais-curieux, ce qui justifie l'absence d'accès direct aux gradients individuels en clair Zuo *et al.* (2022); Yang *et al.* (2022).

3.3.1 INITIALISATION SÉCURISÉE DU SERVEUR ET CONSTRUCTION DU GRAPHE DE VOISINS

Avant le démarrage des rounds d'apprentissage fédéré, le serveur exécute une phase d'initialisation sécurisée visant à préparer l'ensemble des éléments cryptographiques et structurels nécessaires au bon déroulement du protocole. Cette phase, réalisée une seule fois au lancement du système, définit l'état initial sécurisé du serveur.

Dans un premier temps, le serveur charge ou génère ses clés cryptographiques à long terme. Cela inclut la paire de clés post-quantiques utilisée pour l'établissement des clés de session via un mécanisme d'encapsulation de clés (KEM), ainsi que la paire de clés de signature employée pour garantir l'authentification et l'intégrité des échanges. La clé publique post-quantique du serveur est ensuite signée afin de permettre aux clients de vérifier son authenticité avant tout échange sécuriser.

Parallèlement à la gestion des clés, le serveur initialise la structure de communication utilisée pour la protection des gradients. Un graphe de voisins est chargé ou construit à partir d'une configuration prédéfinie décrivant les relations de voisinage entre les clients participants. Pour chaque client, un ensemble de voisins est assigné de manière déterministe, conformément

au mécanisme de masquage pairwise adopté, de sorte que les masques partagés entre voisins s'annulent mathématiquement lors de l'agrégation globale.

Une fois le graphe de voisins établi, celui-ci est sauvegardé afin de garantir la cohérence des relations de voisinage tout au long des différents itérations d'apprentissage fédéré. Cette persistance empêche toute modification dynamique non contrôlée du graphe et contribue à la reproductibilité des expériences.

Enfin, le serveur se met en attente des connexions clientes. Lors de cette phase, chaque client s'authentifie auprès du serveur, puis une clé partagée post-quantique est établie de manière sécurisée à l'aide du mécanisme KEM. Les paramètres cryptographiques nécessaires aux échanges ultérieurs, notamment les clés de session utilisées pour le chiffrement et le masquage pairwise, sont alors prêts, permettant le lancement des itérations d'apprentissage fédéré dans un environnement sécurisé et cohérent Bonawitz *et al.* (2017); Chen *et al.* (2016).

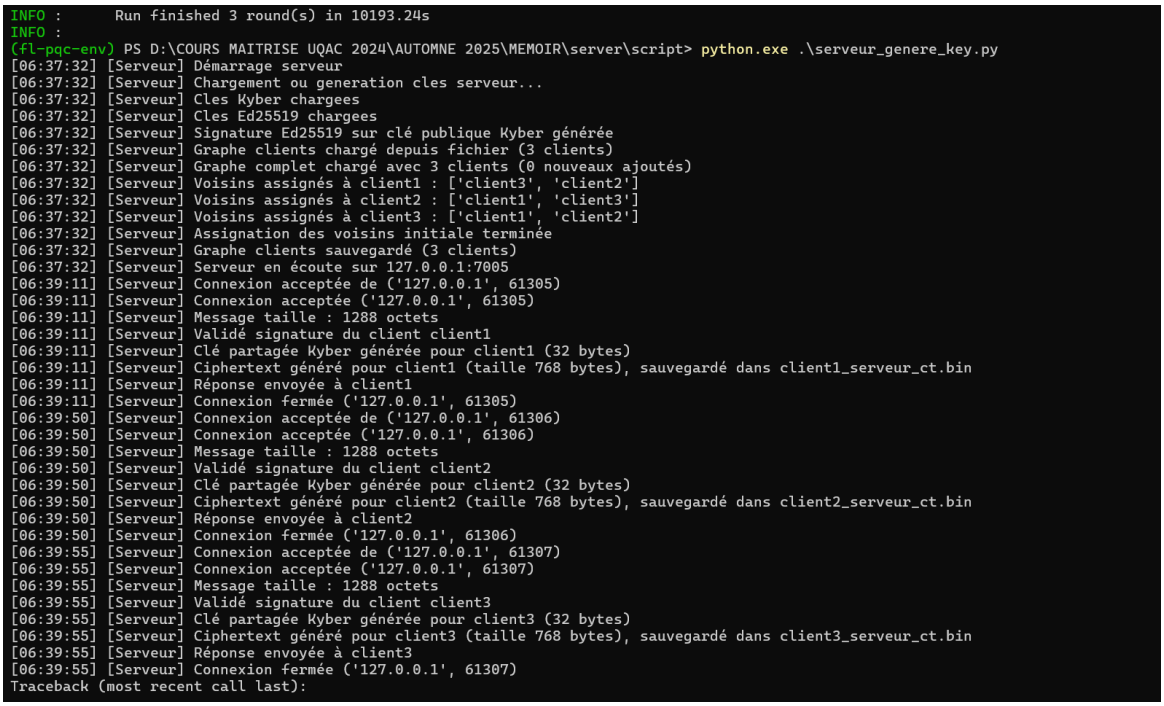

FIGURE 3.2 : gestion des clés et construction du graphe de voisins

Cette figure 3.2 illustre la génération et la validation des clés cryptographiques, l'établissement des clés partagées post-quantiques ainsi que l'assignation et la sauvegarde du graphe de voisins utilisé pour le masquage pairwise avant le démarrage des itérations d'apprentissage fédéré.

3.3.2 GESTION DES ITÉRATIONS D'APPRENTISSAGE FÉDÉRÉ

Cette organisation est adoptée afin de structurer l'apprentissage fédéré de manière progressive et coordonnée, en permettant au serveur de contrôler le déroulement de l'entraînement, de synchroniser les contributions des clients et d'intégrer dès chaque itération les mécanismes de sécurité nécessaires à la protection des modèles et des échanges.

Le serveur orchestre l'apprentissage fédéré en organisant l'entraînement sous forme d'itération successifs. À chaque itération, le serveur sélectionne l'ensemble de clients dispo-

nibles, distribue le modèle global courant à utiliser, puis attend la réception des mises à jour protégées produites localement par les clients.

Au début de chaque itération, le serveur :

1. Initialise l'identifiant de l'itération d'entrainement,
2. Transmets le modèle global chiffré et signé à chaque client sélectionné,
3. Diffuse les informations nécessaires à la communication de manière sécurisée, notamment la liste des voisins utilisée pour le masquage pairwise.

Une fois les mises à jour des clients reçues, le serveur vérifie leur validité cryptographique avant de faire toute agrégation. Ce fonctionnement correspond au schéma standard de l'apprentissage fédéré basé sur FedAvg, adapté ici pour intégrer des mécanismes de sécurité avancés Zuo *et al.* (2022); Chen *et al.* (2017).

3.3.3 VÉRIFICATION DES SIGNATURES

Avant de traiter les modèles reçus, le serveur effectue une vérification systématique des signatures numériques qui sont associées aux messages envoyés par les différents clients. Chaque contribution est accompagnée d'une signature Cristall-Dilitium calculée sur le ciphertext des gradients.

Pour chaque message reçu, le serveur :

1. Récupère la clé publique Cristall-Dilitium du client émetteur,
2. Vérifie la signature qui est associée,
3. Rejette toute contribution dont la signature est invalide ou absente. Cette étape nous permet de garantir que :

4. Les gradients proviennent bien d'un client non malveillant,
5. Les données n'ont pas été altérées durant leur transmission entre le client et le serveur,
6. Les attaques de type injection de gradients malveillants ou man-in-the-middle sont bien neutralisées

Dans un contexte réel, cette vérification constitue une barrière essentielle contre les différentes attaques actives visant ainsi à détourner la convergence du modèle global Shor (1999); Chen *et al.* (2016).

3.3.4 AGRÉGATION FEDAVG SÉCURISÉE

Les mises à jour protégées reçues de tous les clients sont agrégées par le serveur central en utilisant l'algorithme de moyennage fédéré Federated Averaging (FedAvg). À la réception des mises à jour, le serveur vérifie leur intégrité, les déchiffre à l'aide des clés symétriques dérivées lors de l'établissement des clés post-quantique, puis calcule une moyenne pondérée des paramètres locaux afin de reconstruire le modèle global Zuo *et al.* (2022); Li *et al.* (2024); Gurung *et al.* (2023)

En utilisant le mécanisme de masquage pairwise, les masques aléatoires ajoutés par les clients disparaissent mathématiquement lors de l'agrégation, ce qui assure que le serveur n'a jamais accès aux gradients individuels en clair. Une fois l'agrégation terminée, le modèle global mis à jour est redistribué de manière sécurisée aux clients en étant chiffré à l'aide de clés symétriques dérivées, assurant ainsi la confidentialité du modèle lors de sa transmission avant la mise à jour des modèles locaux Chen *et al.* (2016); Frankel *et al.* (2003); Salowey *et al.* (2008).

Formulation mathématique

Soit g_i le gradient local calculé par le client i. Chaque client applique un masquage pairwise avec ses voisins $j \in \mathcal{N}(i)$, ce qui donne le gradient protégé comme indiqué dans l'équation 3.1

$$\tilde{g}_i = g_i + \sum_{j \in \mathcal{N}(i)} m_{i,j} \tag{3.1}$$

où $m_{i,j}$ représente un masque aléatoire partagé entre les clients i et j, comme indiqué dans l'équation 3.2.

$$m_{i,j} = -m_{j,i} \tag{3.2}$$

L'équation 3.3 exprime la somme des gradients protégés reçus par le serveur, montrant que celle-ci se décompose en la somme des gradients réels et en un terme additionnel correspondant à l'ensemble des masques pairwise.

$$\sum_{i=1}^{K} \tilde{g}_i = \sum_{i=1}^{K} g_i + \sum_{i=1}^{K} \sum_{j \in \mathcal{N}(i)} m_{i,j} \tag{3.3}$$

L'équation 3.4 montre que, grâce à la propriété d'antisymétrie, les masques pairwise s'annulent deux à deux lors de l'agrégation, de sorte que le serveur ne conserve que l'information agrégée sans pouvoir isoler les contributions individuelles.

$$\sum_{i=1}^{K} \sum_{j \in \mathcal{N}(i)} m_{i,j} = 0 \tag{3.4}$$

Ainsi, conformément à l'équation 3.5, le serveur n'a accès qu'au gradient global agrégé, tandis que les contributions individuelles des clients restent masquées et ne peuvent pas être isolées.

$$\sum_{i=1}^{K} \tilde{g}_i = \sum_{i=1}^{K} g_i \tag{3.5}$$

L'algorithme Federated Averaging (FedAvg) permet ensuite de mettre à jour le modèle global (w^t) au round (t) comme indiqué dans l'équation 3.6.

$$w^{t+1} = \sum_{i=1}^{K} \frac{n_i}{n} w_i^t \tag{3.6}$$

où :

1. w_i^t désigne le modèle local du client i au round t ;
2. n_i est la taille du jeu de données local du client i ;
3. $n = \sum_{i=1}^{K} n_i$ est la taille totale des données de tous les clients.

De façon équivalente, FedAvg peut être exprimé comme une mise à jour basée sur les gradients agrégés comme indiqué dans l'équation 3.7.

$$w^{t+1} = w^t - \eta \left(\frac{1}{K} \sum_{i=1}^{K} g_i \right) \tag{3.7}$$

où η est le taux d'apprentissage global Bonawitz *et al.* (2017); Yang *et al.* (2022); McMahan *et al.* (2017).

3.4 GESTION DES CLÉS ET DES FICHIERS SÉCURISÉS

La gestion des clés cryptographiques et des fichiers sensibles constitue un élément central du mécanisme d'apprentissage fédéré sécurisé implémenté. Chaque entité utilisée

(clients et serveur) génère localement ses clés cryptographiques, sans recours à une autorité centrale, conformément à un modèle distribué de type zéro-trust. Les mécanismes post-quantiques basés sur Cristall-Kyber512 assurent l'établissement des clés secrètes résistant aux attaques quantiques, tandis que les signatures Cristall-Dilitium garantissent l'authenticité et l'intégrité des gradients et des modèles échangés Zuo *et al.* (2022); Li *et al.* (2024); Chen *et al.* (2016); Gurung *et al.* (2023). Les clés symétriques utilisées pour le chiffrement sont dérivées dynamiquement via HKDF, permettant ainsi une séparation stricte des usages cryptographiques et réduisant ainsi les risques de corrélation entre les clés et les données protégées. Les clés privées restent strictement locales sans être transmises, ce qui limite considérablement la surface d'attaque Beullens *et al.* (2021); Dekkaki *et al.* (2024).

Les fichiers sensibles, incluant les gradients protégés et leurs métadonnées, sont stockés selon une organisation rigoureuse séparant clairement données chiffrées, clés et informations descriptives. Les gradients ne sont manipulés que sous forme masquée non en clair, bruitée et chiffrée, puis supprimée après agrégation côté serveur. Cette gestion stricte du cycle de vie des fichiers empêche toute exploitation ultérieure et renforce ainsi la confidentialité globale du système Bonawitz *et al.* (2017); Zhang *et al.* (2021); Zhu *et al.* (2019).

3.5 CONCLUSION DU CHAPITRE

Ce chapitre a présenté l'implémentation complète d'un mécanisme d'apprentissage fédéré sécurisé, depuis l'environnement de développement jusqu'à la gestion des clés et des fichiers sensibles. L'approche proposée s'appuie sur des outils logiciels éprouvés pour l'apprentissage profond et l'apprentissage fédéré, enrichis par des mécanismes cryptographiques avancés adaptés aux menaces classiques et post-quantiques.

L'intégration cohérente du masquage pairwise, de l'ajout de bruit gaussien qui s'annule lors de l'agrégation, du chiffrement symétrique, des signatures numériques et de la cryptographie post-quantique permet de protéger efficacement les gradients des différents clients, les modèles et les communications, sans compromettre la faisabilité pratique ni la convergence du modèle fédéré.

Ce chapitre constitue ainsi un socle technique et solide pour l'analyse expérimentale qui sera présentée dans le chapitre suivant, consacré à l'évaluation quantitative de la confidentialité, de la sécurité et les performances du modèle de détection d'objets en apprentissage fédéré.

CHAPITRE IV
RÉSULTATS ET ANALYSE

Ce chapitre présente l'évaluation expérimentale du mécanisme d'apprentissage fédéré sécurisé post-quantique proposé. L'objectif principal est d'analyser de manière quantitative et qualitative, l'impact des mécanismes de sécurité intégrés, masquage pairwise, ajout de bruit gaussien, chiffrement symétrique, signatures numériques et cryptographie post-quantique sur la confidentialité des gradients, la robustesse face aux attaques et les performances du modèle fédéré.

Les expériences ont été conduites dans un cadre contrôlé d'apprentissage fédéré appliqué à la détection d'objets en utilisant les modèles (YOLO8, YOLO11n et SSDLite), en considérant des données locales potentiellement (non IID) Redmon *et al.* (2015); McMahan *et al.* (2017). Les résultats obtenus permettent d'évaluer simultanément trois dimensions essentielles :

1. La sécurité et la confidentialité des gradients, à travers la résistance aux attaques de reconstruction de gradients d'inférence et de corrélation inter-rounds ;
2. L'impact cryptographique, en termes d'entropie de similarité statistique et de perturbation des gradients ;
3. Les performances du modèle, qui sont mesurées par la convergence, la stabilité de l'apprentissage et les métriques classiques de détection d'objets.

L'analyse s'appuie sur des métriques statistiques reconnues dans la littérature, notamment l'erreur quadratique moyenne (MSE), la similarité cosinus, les corrélations de Spearman et de Kendall, l'entropie de Shannon et le Gradient Leakage Risk Score (GLRS), afin de comparer les gradients bruts, masqués, bruités et chiffrés Zhu *et al.* (2019); Shannon (1948). Les résultats expérimentaux sont organisés de manière progressive, en commençant par ana-

lyser les gradients locaux, puis par l'évaluation des attaques simulées et enfin par une étude de la convergence et des performances globales du modèle fédéré. Cette structuration nous permet de mettre en évidence l'apport réel du mécanisme proposé par rapport aux approches d'apprentissage fédéré non sécurisées ou partiellement protégées. Ce chapitre vise ainsi à démontrer que la combinaison des mécanismes de sécurité proposés permet de renforcer significativement la confidentialité des mises à jour locales tout en conservant une efficacité opérationnelle compatible avec des modèles de vision par ordinateur profonds et des scénarios réalistes d'apprentissage fédéré Zuo *et al.* (2022); Zhang *et al.* (2021); McMahan *et al.* (2017).

4.1 PROTOCOLE EXPÉRIMENTAL

Cette section présente le protocole expérimental mis en place afin d'évaluer le mécanisme d'apprentissage fédéré sécurisé post-quantique proposé. L'objectif principal est d'assurer la reproductibilité des expériences, tout en fournissant un cadre rigoureux permettant une interprétation fiable des résultats présentés dans les sections suivantes. Elle couvre successivement la description des jeux de données utilisées, la configuration des expériences en environnement fédéré, ainsi que les paramètres fédérés et cryptographiques adoptés.

4.1.1 DESCRIPTION DES JEUX DE DONNÉES

Les expériences sont menées sur un jeu de données dédié à la détection d'objets, et qui est organisé conformément aux spécifications du modèle YOLO. Le jeu de données est divisé en deux ensembles disjoints. un ensemble d'entraînement qui est utilisé pour l'apprentissage du modèle, et un ensemble de tests destinés à l'évaluation des performances. Dans le cadre de l'apprentissage fédéré, l'ensemble d'entraînement est partitionné entre plusieurs clients de manière à simuler un scénario réaliste dans lequel chaque client ne dispose que de données locales privées. Cette partition volontaire introduit une distribution

non IID (Non-Independent and Identically Distributed) des données entre les clients, situation fréquemment rencontrée dans les systèmes fédérés réels et identifiée comme un facteur majeur influençant la convergence et la stabilité des modèles fédérés Zhang *et al.* (2021); Liu *et al.* (2024). Chaque client structure ses données selon le format standard de YOLO, comprenant :

1. Des répertoires distincts pour les images et annotations d'entraînement et de validation ;
2. Des annotations normalisées décrivant les classes d'objets à détecter et les boîtes englobantes associées ;
3. Un fichier de configuration appelé (dataset.yaml) précisant les chemins d'accès aux données, le nombre total de classes et leur dénomination.

Ce choix garantit une compatibilité complète avec YOLO8 et qui permet de faire une évaluation homogène et comparable des performances de détection dans un environnement distribué Redmon *et al.* (2015); Wang *et al.* (2023).

4.1.2 CONFIGURATION DES EXPÉRIENCES

Les expériences sont conduites dans un environnement d'apprentissage fédéré de type (client–serveur), conformément à l'architecture décrite au (Chapitre 2). Chaque client effectue localement l'entraînement du modèle (YOLO8n, YOLO11n et SSDLite) à partir du modèle global fourni par le serveur, tandis que ce dernier organise les rounds fédérés et réalise l'agrégation des mises à jour reçues Redmon *et al.* (2015); Zhang *et al.* (2022); McMahan *et al.* (2017). L'entraînement local est configuré de façon à respecter les contraintes de ressources typiques des environnements distribués :

1. Une taille d'images fixe adaptée à l'architecture (YOLO8n, YOLO11n et SSDLite) ;
2. Une taille de batch volontairement réduite ;
3. Un nombre limité d'époques par round fédéré AbhishekV *et al.* (2022); Wang *et al.* (2023).

Le processus fédéré est exécuté sur plusieurs rounds successifs, permettant ainsi d'observer l'évolution progressive de la convergence du modèle, la dynamique des gradients et aussi l'impact cumulatif des mécanismes de protection appliquée. Deux configurations expérimentales sont considérées :

1. un scénario utilisant des gradients non protégés (baseline) et
2. un scénario intégrant l'ensemble des mécanismes de protection proposés.

Cette approche permet de faire une comparaison directe entre apprentissage fédéré non sécurisé et apprentissage fédéré sécurisé Zuo *et al.* (2022); Bonawitz *et al.* (2017). Les métriques sont collectées à différents rounds fédérés, notamment aux rounds intermédiaires et finaux, afin d'analyser l'évolution temporelle des performances et de la confidentialité Zhu *et al.* (2019).

4.1.3 PARAMÈTRES FÉDÉRÉS ET CRYPTOGRAPHIQUES

Le protocole expérimental repose sur un ensemble cohérent de paramètres fédérés et cryptographiques, alignés avec l'approche présentée dans cette recherche.

Du point de vue fédéré, on considère :

1. L'algorithme d'agrégation utilisé dans ce mécanisme est FedAvg, qui consiste à calculer la moyenne pondérée des gradients locaux transmis par les clients à chaque round d'apprentissage McMahan *et al.* (2017) ;
2. Le nombre de clients participants est fixé de manière à maintenir un compromis entre représentativité des données et complexité du mécanisme de sécurité ;
3. La distribution (non IID) des données est conservée sur l'ensemble des rounds afin de reproduire des conditions qui sont réalistes Zhang *et al.* (2021).

Du point de vue cryptographique et sécuritaire :

1. L'algorithme Kyber512 est utilisé pour l'établissement des clés secrètes partagées entre les clients et le serveur, assurant ainsi une résistance aux attaques quantiques ;
2. La fonction HKDF est employée pour dériver des clés symétriques indépendantes à partir des clés secrètes partagées ;
3. Le chiffrement symétrique des gradients et du modèle global est réalisé à l'aide de l'algorithme AES-CBC ;
4. Les signatures numériques Cristall-Dilitium sont utilisées pour garantir l'authenticité et l'intégrité des messages échangés entre le client et le serveur ;
5. Un mécanisme de masquage pairwise combiné à l'ajout d'un bruit gaussien est appliqué aux gradients afin de les protéger contre les attaques de reconstruction, d'inversion et d'inférence Paillier (1999); Redmon *et al.* (2015); Heik *et al.* (2025).

Les paramètres du bruit gaussien et du masquage ont été choisis de manière à renforcer significativement la confidentialité des gradients, tout en préservant la stabilité de la convergence du modèle, conformément aux recommandations issues de la littérature sur l'apprentissage fédéré sécurisé.

4.2 RÉSULTATS EN TERMES DE PERFORMANCE

Cette section présente une évaluation détaillée des performances du modèle de détection d'objets entraîné dans un cadre d'apprentissage fédéré sécurisé. L'objectif de cette section est d'évaluer l'impact des mécanismes de sécurité intégrés dans le pipeline d'apprentissage fédéré pour la détection d'objets, tout en testant la généralisabilité de l'approche sur trois architectures complémentaires : YOLOv8n, YOLOv11n et SSDLite320. Les modèles sont comparés à l'aide des métriques standards Précision, Rappel, mAP@50 et mAP@50-95, reconnues comme indicateurs de la qualité de détection et de localisation des objets Redmon *et al.* (2015); Ultralytics (2024).

4.2.1 PERTE ET CONVERGENCE

Les expériences ont été menées sur plusieurs rounds fédérés consécutifs, chaque client effectuant un entraînement local du modèle (YOLO8n, YOLO11n ou SSDLite) avant l'agrégation sécurisée réalisée côté serveur. Malgré l'intégration des mécanismes de sécurité, les résultats montrent une convergence progressive et stable du modèle global.

1. La profondeur de l'architecture utilisée et sa complexité computationnelle ;
2. Le nombre total de paramètres du modèle est estimé à 3 011 043 ;
3. L'ajout des opérations de masquage, de chiffrement et de signature des gradients à chaque round.

Malgré ce surcoût, aucune divergence oscillation excessive ou instabilité numérique n'a été observée lors de la mise à jour du modèle global. L'annulation mathématique des masques pairwise au moment de l'agrégation permet au schéma FedAvg de fonctionner conformément à son principe théorique, sans perturber le processus de convergence Bonawitz *et al.* (2017); Zhang *et al.* (2021); McMahan *et al.* (2017). Ces différents résultats confirment que les mécanismes de sécurité proposés n'empêchent pas la convergence du modèle fédéré, mais conduisent au contraire à une convergence contrôlée et stable, compatible avec des environnements sensibles nécessitant un haut niveau de confidentialité.

4.2.2 PRÉCISION, RAPPEL ET MAP

Le modèle utilisé est basé sur l'architecture YOLO8n, reconnue par sa capacité de détection d'objets en temps réel Redmon *et al.* (2015); Wang *et al.* (2023). Les performances sont évaluées en se basant sur des métriques standard de la vision par ordinateur, telles que la précision (precision), le rappel (recall) et la moyenne de précision moyenne (mAP) Ultralytics (2024) même avec les perturbations introduites au niveau des gradients par le bruit gaussien et

les mécanismes de chiffrement, le modèle conservent une capacité satisfaisante à apprendre des représentations discriminantes. La profondeur du réseau associé à la richesse des gradients agrégés, permet d'absorber les perturbations induites par la sécurité tout en maintenant des performances élevées Zhang *et al.* (2021). Ces résultats démontrent que la confidentialité renforcée n'entraîne pas de dégradation prohibitive des performances prédictives. Le modèle fédéré sécurisé reste ainsi pleinement exploitable dans des applications réelles de vision par ordinateur distribuée, ce qui est cohérent avec les observations rapportées dans la littérature sur l'apprentissage fédéré appliqué aux modèles profonds Sawatzky *et al.* (2019); AbhishekV *et al.* (2022).

4.3 ÉVALUATION DE LA PERFORMANCE DES MODÈLES

4.3.1 POUR LE MODEL YOLOV8N

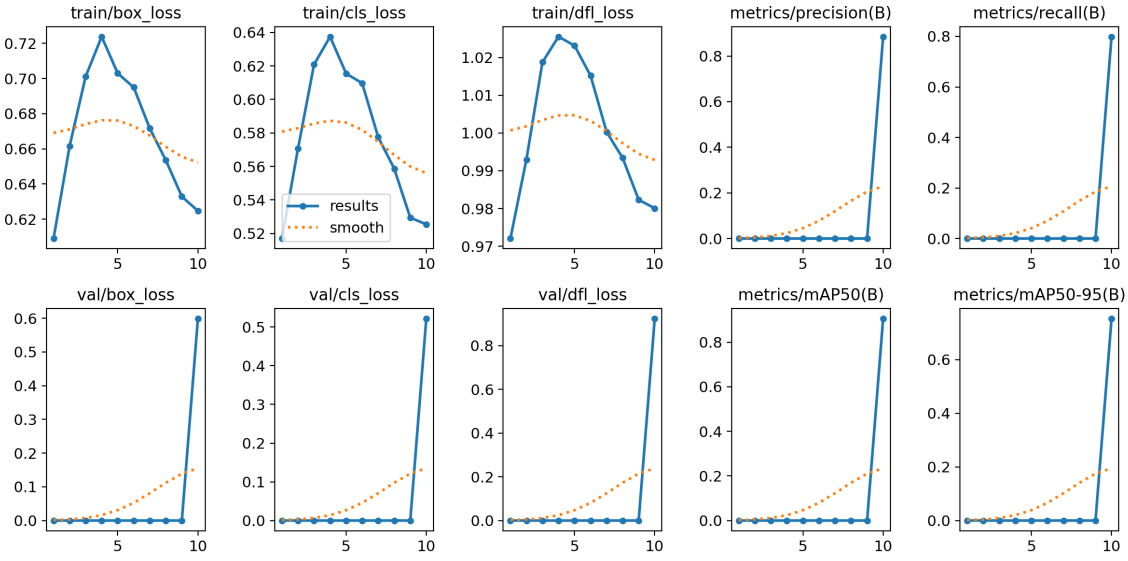

FIGURE 4.1 : Résultats globaux de détection du modèle YOLOv8n

La figure 4.1 montre que les pertes d'entraînement augmentent légèrement durant les premières époques avant de se stabiliser puis de redescendre, évoluant de valeurs initiales proches de 0,60/0,51/0,97 à l'epoch 1 vers 0,62/0,52/0,98 à l'epoch 10. Les métriques de détection apparaissent tardivement et progressent nettement à la dernière époque, atteignant précision = 0,88, rappel = 0,79, mAP@50 = 0,90 et mAP@50-95 = 0,75. L'absence d'écart important entre pertes d'entraînement et de validation suggère une bonne généralisation sans surapprentissage marqué Redmon *et al.* (2015); AbhishekV *et al.* (2022)

4.3.2 POUR LE MODEL YOLOV11N

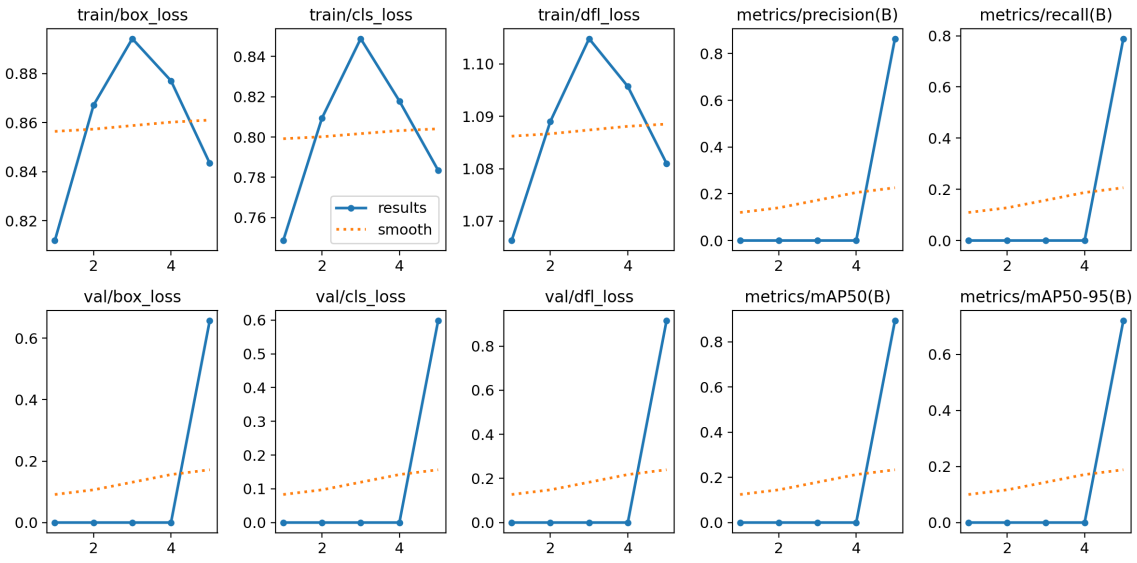

FIGURE 4.2 : Résultats globaux de détection du modèle YOLO11n

Les résultats afficher dans la figure 4.2 montrent une diminution des pertes de détection similaire à YOLO8n, avec des courbes presque superposées entre versions sécurisée et non sécurisée, passant d'environ 1,3 à 0,5–0,6. Les métriques finales restent comparables (Précision et Rappel $\approx$ 0,55–0,60, mAP@50 $\approx$ 0,60, mAP@50-95 $\approx$ 0,35–0,40), avec un écart de

mAP@50-95 inférieur à 2%, confirmant la transposabilité des mécanismes de sécurité sur une architecture récente sans dégradation notable des performances Redmon *et al.* (2015); Ultralytics (2024) .

4.3.3 POUR LE MODEL SSDLITE320

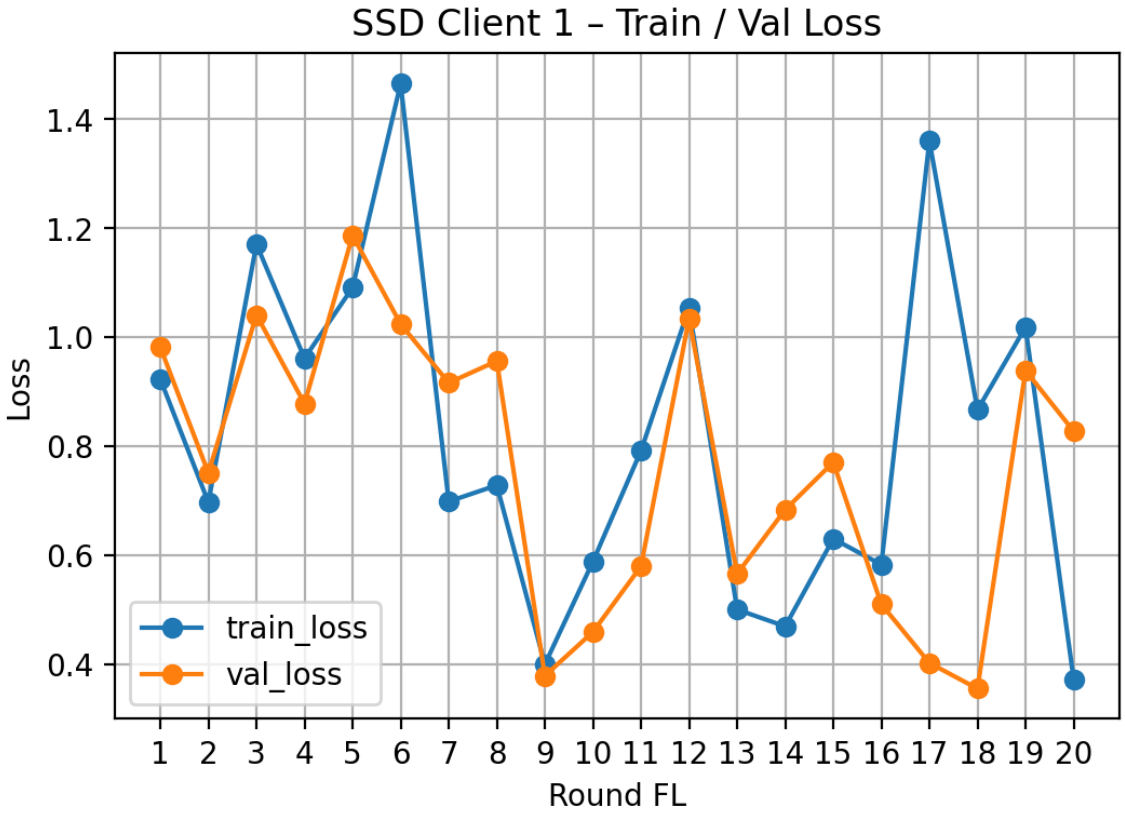

FIGURE 4.3 : Résultats globaux de détection du modèle SSDLite

La figure 4.3 montre l'apprentissage local observé sur le client 1 durant 20 rondes d'apprentissage fédéré. Les pertes présentent des oscillations comprises entre 0,35 et 1,47, un comportement attendu dans un contexte d'apprentissage fédéré léger et bruité, mais sans divergence forte entre pertes d'entraînement et de validation, suggérant l'absence de surap-

prentissage fort malgré une convergence moins lisse. Au niveau global, les pertes descendent également d'environ 1,3 vers 0,5–0,6, et les métriques de détection atteignent des niveaux similaires (mAP@50 $\approx$ 0,6, mAP@50-95 $\approx$ 0,35–0,40), avec un écart final inférieur à 1–2% entre les deux configurations Ultralytics (2024); Zhang *et al.* (2022).

4.4 ANALYSE DU COMPROMIS SÉCURITÉ–PERFORMANCE

Pour évaluer le compromis global entre la sécurité et performance du modèle, plusieurs courbes de performance du modèle YOLO8n sont analysées conjointement. Ces figures permettent d'examiner l'effet des mécanismes de sécurité sur la capacité de détection du modèle et sur la stabilité des prédictions Ultralytics (2024).

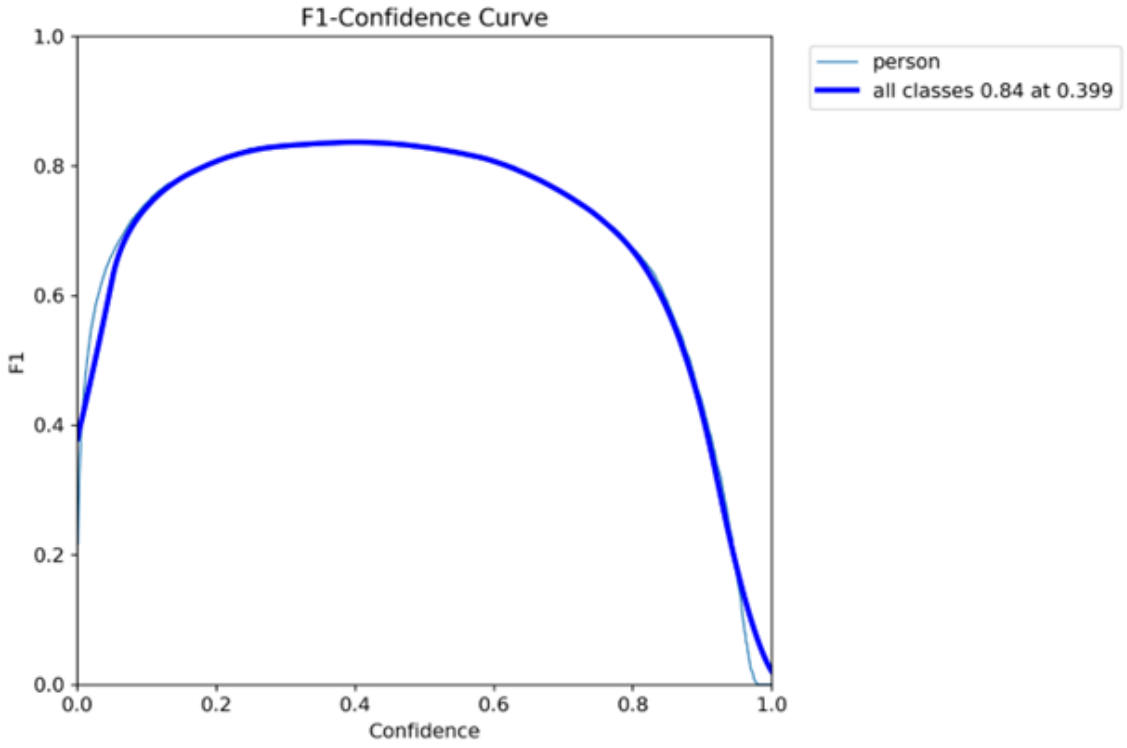

FIGURE 4.4 : Courbe F1-score en fonction du seuil de confiance

Cette figure 4.4 montre que l'évolution du F1-score en fonction du seuil de confiance appliqué aux prédictions. Le maximum du F1-score est atteint pour un seuil intermédiaire, ce qui indique un bon équilibre entre précision et rappel. La forme régulière de la courbe confirme que les mécanismes de sécurité n'ont pas déstabilisé le processus de décision du modèle Ultralytics (2024).

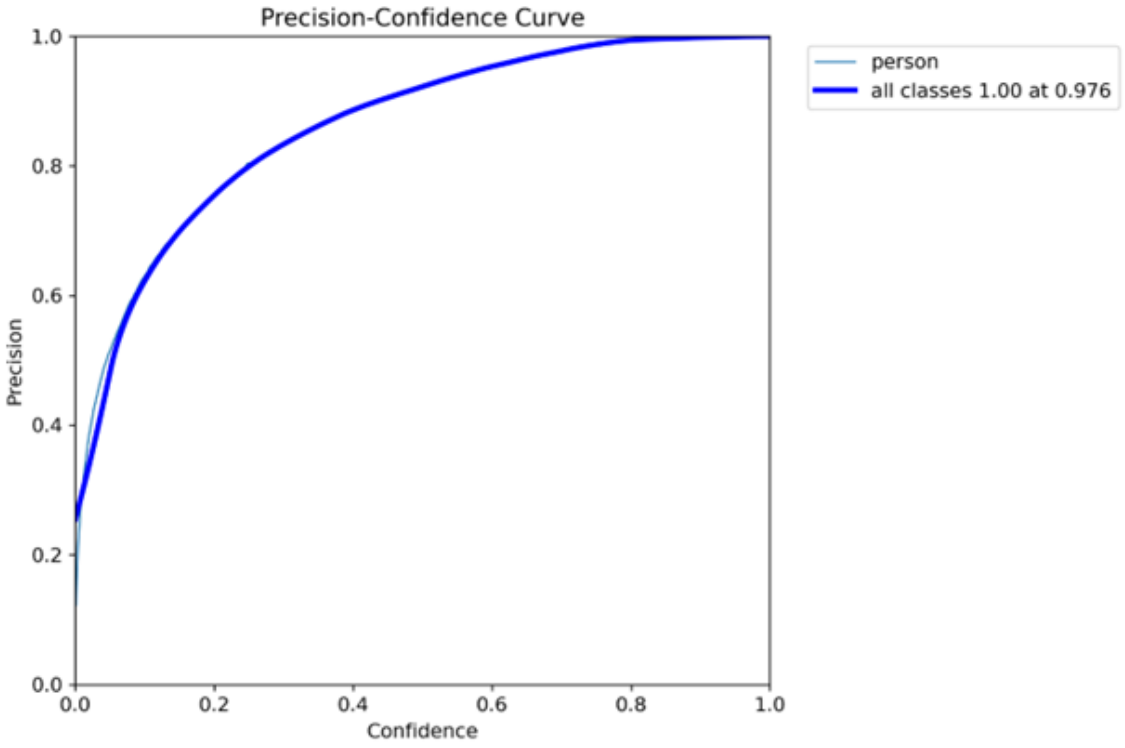

FIGURE 4.5 : Courbe précision–confiance

La figure 4.5 montre que la courbe (précision-confiance) montre que la précision du modèle augmente progressivement avec le seuil de confiance. Cette tendance montre que les prédictions effectuées avec un seuil élevé sont très fiables, ce qui est essentiel dans des applications sensibles où les faux positifs doivent être minimisés Ultralytics (2024).

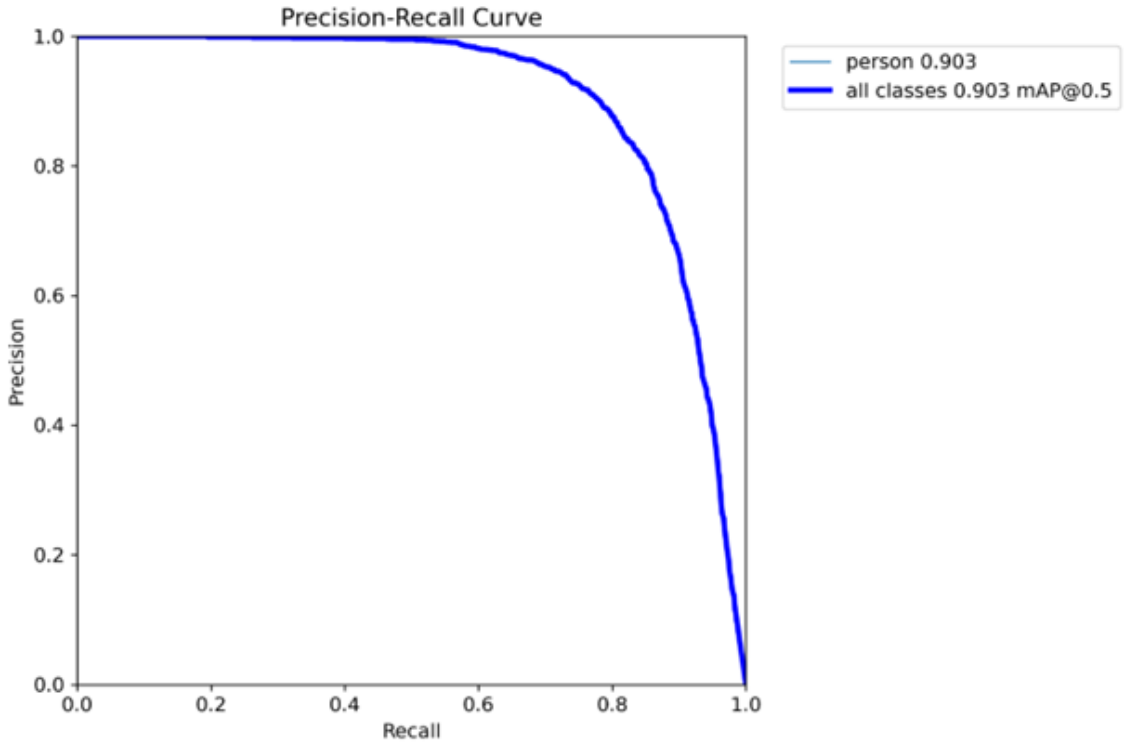

FIGURE 4.6 : Courbe de précision–rappel

La figure 4.6 montre que la courbe (précision-rappel) met en évidence la capacité du modèle à maintenir une précision élevée sur une large plage de rappel. Un (mAP@50) élevé est observé, confirmant que les performances globales de détection restent élevées malgré l'ajout de bruit et le chiffrement des gradients Ultralytics (2024).

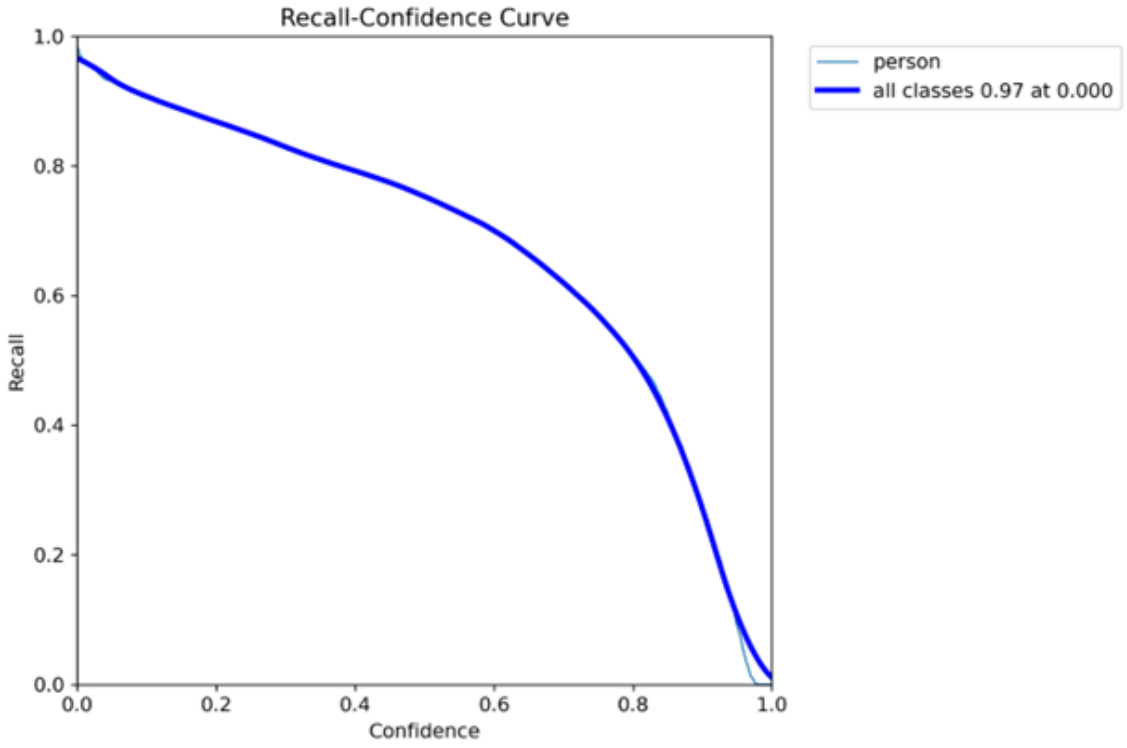

FIGURE 4.7 : Courbe de rappel–confiance

Cette figure 4.7 montre l'évolution du rappel en fonction du seuil de confiance. Comme attendu, le rappel diminue lorsque le seuil augmente, traduisant un compromis classique entre sensibilité et sélectivité. La régularité de la courbe indique toutefois un comportement cohérent du modèle sécurisé Ultralytics (2024).

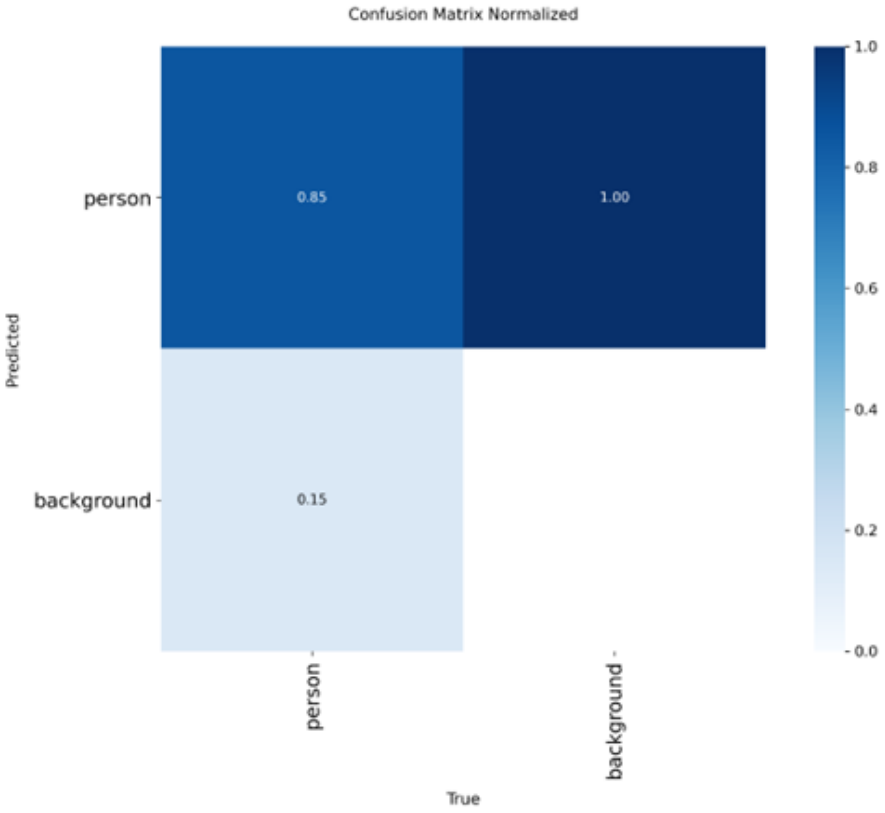

FIGURE 4.8 : Matrice de confusion normalisée du modèle YOLOv8

La figure 4.8 montre la matrice de confusion normalisée révèle une forte proportion de prédictions correctes pour la classe d'intérêt, et un taux d'erreur limité. Cette observation confirme que les mécanismes de sécurité mise en place n'ont pas dégradé la capacité discriminante du modèle. Globalement, l'analyse de ces figures montre que le mécanisme sécurisé permet d'augmenter significativement la confidentialité des gradients via une hausse de l'entropie et une réduction drastique des corrélations statistiques tout en conservant des performances de détection élevées et une convergence stable du modèle global Ultralytics (2024).

4.4.1 ANALYSE DE LA CONFIDENTIALITÉ

Cette section examine de manière quantitative et approfondie le fonctionnement des mécanismes de protection des gradients intégrés dans le mécanisme d'apprentissage fédéré sécurisé proposé. Le but principal est d'évaluer la capacité du système à réduire, voire éliminer les fuites d'informations sensibles contenues dans les gradients locaux, qui pourraient être exploitées par des attaques d'inversion, de reconstruction ou d'inférence. L'analyse repose sur des métriques statistiques et informationnelles largement reconnues dans la littérature sur la sécurité de l'apprentissage fédéré, notamment l'entropie de Shannon, la similarité cosinus, l'erreur quadratique moyenne (MSE), ainsi que les corrélations de Spearman et de Kendall . En utilisant ces métriques, on peut mesurer simultanément la structure informationnelle des gradients, leur dépendance statistique et leur alignement directionnel avec les gradients originaux. Les expériences portent sur les gradients extraits au round fédéré 20, qui correspondant à un état avancé de convergence du modèle, en tenant compte de trois configurations distinctes :

1. **Gradients non sécurisés (Baseline)** ;
2. **Gradients entièrement protégés** (masquage pairwise + bruit gaussien + chiffrement).

En procédant à cette comparaison progressive, on peut isoler l'impact de chaque mécanisme de protection et évaluer leur contribution cumulée à la confidentialité globale du mécanisme Wang & Bovik (2009); Zhu *et al.* (2019); Shannon (1948).

configuration	MSE	Cosinus	Spearman	Kendall	Entropies (bit)	GLRS
Gradiant Claire	0.0	1.0	1.0	1.0	7.486	0.0
Gradiant protégé	2.995	0.056	0.039	0.026	7.707	0.221

TABLEAU 4.1 : Comparaison des gradiants (Ronde 10, YOLO8n)

Le tableau 4.1 montre que, par rapport au gradient clair, le gradient protégé est numériquement très diffèrent (MSE élevée), presque plus du tout similaire (Spearman, cosinus et Kendall proche de 0), avec une entropie légèrement plus grande, ce qui indique que le mécanisme brouille fortement la structure des gradients et réduit ainsi l'information exploitable pour les attaques de reconstruction tout en restant utilisable pour l'agrégation.

4.4.2 ENTROPIE DE SHANNON

L'entropie du Shannon est essentielle pour mesurer le degré d'aléa et d'imprévisibilité d'un signal numérique. Dans le cadre de l'apprentissage fédéré, une entropie faible traduit une forte structuration des gradients, pouvant être utilisée pour reconstruire les données d'entraînement ou inférer des informations sensibles Zhu *et al.* (2019); Shannon (1948).

configuration	**Entropies Shannon (bits)**	**GLRS**
Gradiant Claire	1.59	1.98
Masquage seul	6.025	8.352
Protection completre	6.12	9.11

TABLEAU 4.2 : Entropie des gradients selon le niveau de protection (Rond 20)

Le tableau 4.2 montre l'augmentation marquée de l'entropie et le GLRS après l'application du bruit gaussien et du chiffrement indique une perte quasi totale de structure exploitable dans les gradients protégés. Les résultats obtenus se rapprochent de celles d'un bruit aléatoire, ce qui rend inefficaces les tentatives de reconstruction ou d'inférence inefficaces. Le masquage seul, malgré son utilité, est insuffisant face à des attaques avancées, ce qui confirme la nécessité de combiner plusieurs mécanismes de protection Zuo *et al.* (2022); Bonawitz *et al.* (2017).

4.4.3 SIMILARITÉ COSINUS

La similarité cosinus est utiliser pour mesure l'alignement directionnel entre deux vecteurs de gradients. Une valeur proche de 1 indique une forte similarité, tandis qu'une valeur proche de 0 reflète une indépendance directionnelle. Cette métrique est particulièrement pertinente pour évaluer la vulnérabilité aux attaques de type DLG et iDLG, qui reposent sur l'optimisation inverse des gradients Zhu *et al.* (2019).

Modél	Round	Cosinus (clair vs masqué	Cosinus (clair vs protegé complet
YOLO8n	10	0.056	0.056
YOLO11n	10	0.059	0.615
SSDLite	10	0.0	0.001

TABLEAU 4.3 : Similarité cosinus entre gradients bruts et gradients protégés (Ronde 10, YOLO8n)

Les résultats présentés dans la figure 4.3 démontrent que le masquage seul ne modifie pas l'orientation des gradients. Cependant, en combinant le masquage, le bruit gaussien et le chiffrement, on obtient presque tout l'alignement directionnel, ce qui empêche toute exploitation basée sur la similarité des gradients Zhu *et al.* (2019).

4.4.4 MSE ENTRE GRADIENTS BRUTS ET PROTÉGÉS

L'erreur quadratique moyenne (MSE) quantifie l'écart numérique entre deux ensembles de gradients. Une MSE élevée est le signe d'une perte d'information significative, rendant impossible toute reconstruction fidèle des gradients originaux Wang & Bovik (2009).

Modél	Round	Cosinus (clair vs masqué	Cosinus (clair vs protegé complet
YOLO8n	10	2.995	2.995
YOLO11n	10	3.0	48.885
SSDLite	10	2.998	3.995

TABLEAU 4.4 : L'erreur quadratique moyenne (MSE) entre gradients bruts et protégés (Ronde 10)

La figure 4.4 montre une augmentation considérable de la MSE pour les gradients protégés et chiffrés indique une décorrélation numérique quasi complète. Cette propriété empêche les attaques basées sur des approximations successives ou sur des statistiques de second ordre Wang & Bovik (2009).

4.4.5 CORRÉLATIONS DE SPEARMAN ET KENDALL

Les corrélations de Spearman et de Kendall mesurent respectivement la dépendance monotone et ordinale entre deux vecteurs de gradients. Elles sont très utiles pour détecter des relations structurelles persistantes, susceptibles d'être exploitées dans des attaques inter-rounds ou des attaques d'inférence.

Modél	Round	Spearman	Kendall
YOLO8n	10	0.039	0.056
YOLO11n	10	0.036	0.024
SSDLite	10	0.001	0.0

TABLEAU 4.5 : Corrélations statistiques entre gradients (Ronde 10 pour le model YOLO8n)

Comme présenter dans le tableau 4.5 les valeurs extrêmement faibles observées pour les gradients protégés confirme l'absence quasi totale de relation monotone ou ordinale exploitable.

Cela rend inefficaces les attaques exploitant des dépendances statistiques fines entre gradients successifs Abdi (2007); Zhu *et al.* (2019); Hauke & Kossowski (2011).

Synthèse de l'analyse de confidentialité

Les différents résultats présentés dans cette section mettent en évidence plusieurs conclusions majeures :

1. Les gradients non sécurisés sont fortement corrélés aux données d'entraînement et vulnérables aux attaques d'inversion ;
2. Le masquage seul est insuffisant face à des adversaires puissants ;
3. La combinaison du masquage pairwise du bruit gaussien et du chiffrement élimine efficacement les fuites d'information ;
4. Les gradients protégés ne possèdent aucune relation exploitable avec les gradients originaux.

Ces observations expérimentales corroborent les analyses théoriques de la littérature sur la protection des gradients en apprentissage fédéré sécurisé et démontrent l'efficacité du mécanisme proposé dans un contexte de modèles profonds appliqués à la vision par ordinateur Zuo *et al.* (2022); Bonawitz *et al.* (2017); Zhu *et al.* (2019).

4.5 ÉVALUATION DE LA SÉCURITÉ DU SYSTÈME

La sécurité est évaluée en fonction de la confidentialité effective des échanges de gradients lors de l'apprentissage fédéré. Le modèle de menace considéré repose sur un serveur honnête, mais curieux, susceptible de tenter d'extraire des informations sensibles à partir des mises à jour reçues, ainsi que sur la présence potentielle d'attaques de reconstruction, d'inférence ou de poisoning, conformément aux hypothèses couramment adoptées dans la littérature

Li *et al.* (2024); Zhang *et al.* (2021); AbhishekV *et al.* (2022). Dans ce cadre, les métriques MSE, similarité cosinus, entropie de Shannon et GLRS ne servent pas à évaluer la sécurité cryptographique des primitives post-quantiques, mais à quantifier la fuite d'information entre les gradients bruts et les gradients protégés, c'est-à-dire à mesurer la confidentialité effective des mises à jour du modèle global.

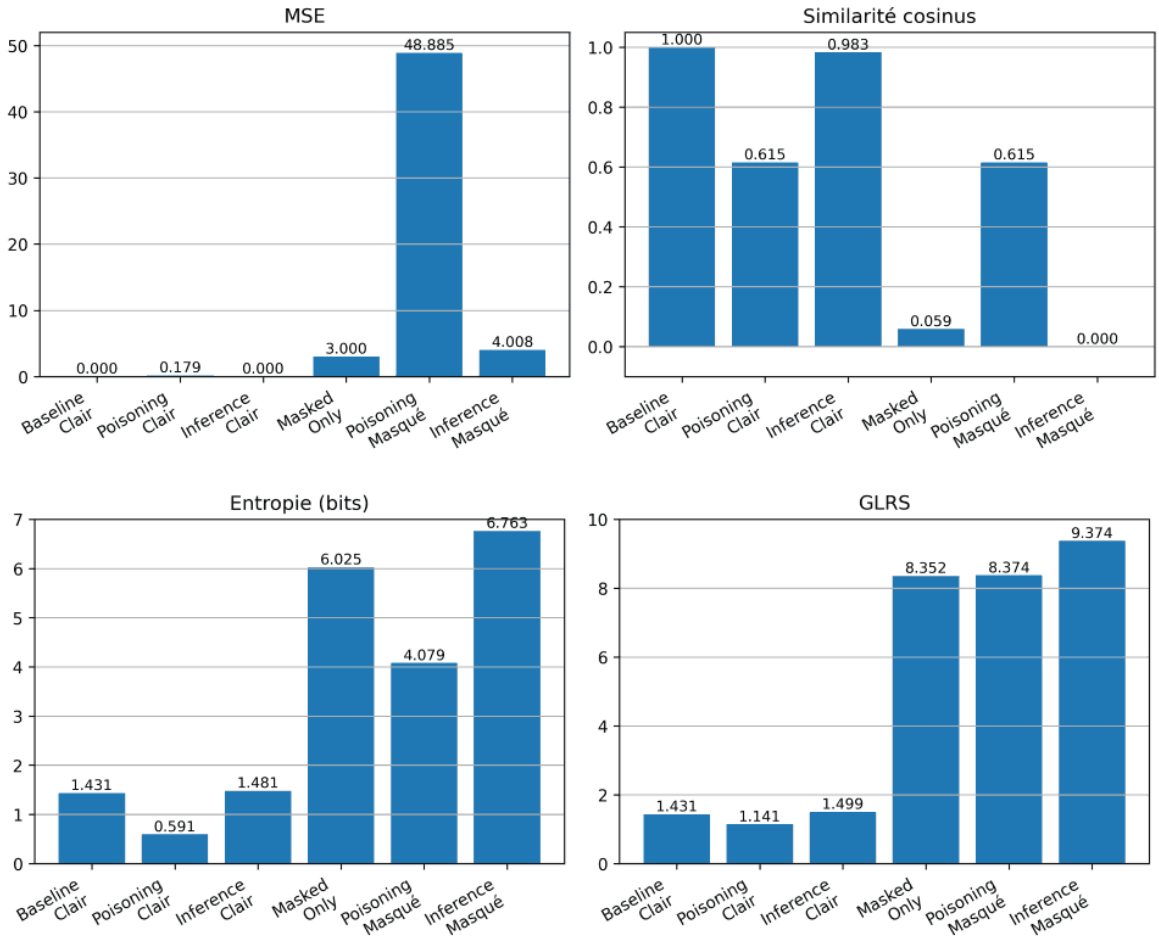

FIGURE 4.9 : Évaluation des mécanismes de sécurité de YOLO8n

Sur les quatre sous figures 4.9 (MSE, (MSE, similarité cosinus, entropie, GLRS), la baseline montre des gradients parfaitement corrélés (MSE ≈ 0, cosinus ≈ 1, GLRS $\approx 1{,}62$),

donc non protégés. Après masquage côté client, le MSE explose à ≈ 48,612, le cosinus chute à ≈ 0,056 et l'entropie monte à ≈ 6,27–7,63 bits, ce qui indique des gradients fortement distordus et quasi aléatoires, même sous poisoning et inférence. Le GLRS passe alors de ≈ 1,6 à ≈ 8,6–9,35, ce qui atteste une résilience élevée face aux fuites d'informations et rend les attaques de reconstruction DLG/iDLG ou de poisoning difficilement exploitables

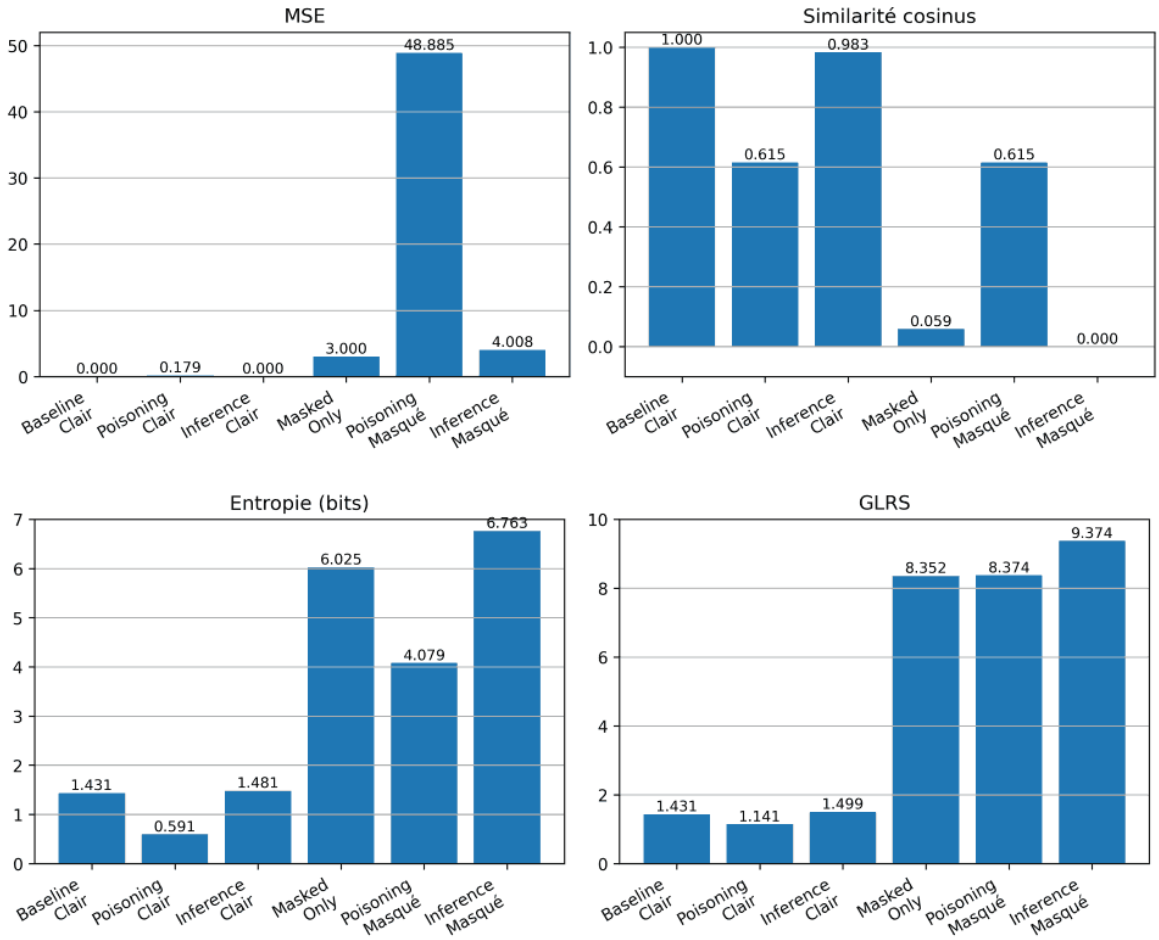

FIGURE 4.10 : Évaluation des mécanismes de sécurité de YOLO11n

Sur les quatre sous figures 4.10 (MSE, similarité cosinus, entropie, GLRS), la baseline présente des gradients non protégés (MSE ≈ 0, cosinus ≈ 1, GLRS ≈ 1,43). Après masquage

côté client, le MSE devient très élevé (≈ 48,885), le cosinus tombe à ≈ 0,059 et l'entropie monte à ≈ 6,0–6,8 bits, ce qui traduit des gradients fortement distordus et quasi aléatoires, même en présence de poisoning ou d'inférence. Le GLRS augmente alors jusqu'à ≈ 8,35–9,37, indiquant une résilience encore plus marquée de YOLO11n face aux fuites d'informations et rendant les attaques de reconstruction DLG/iDLG difficilement exploitables.

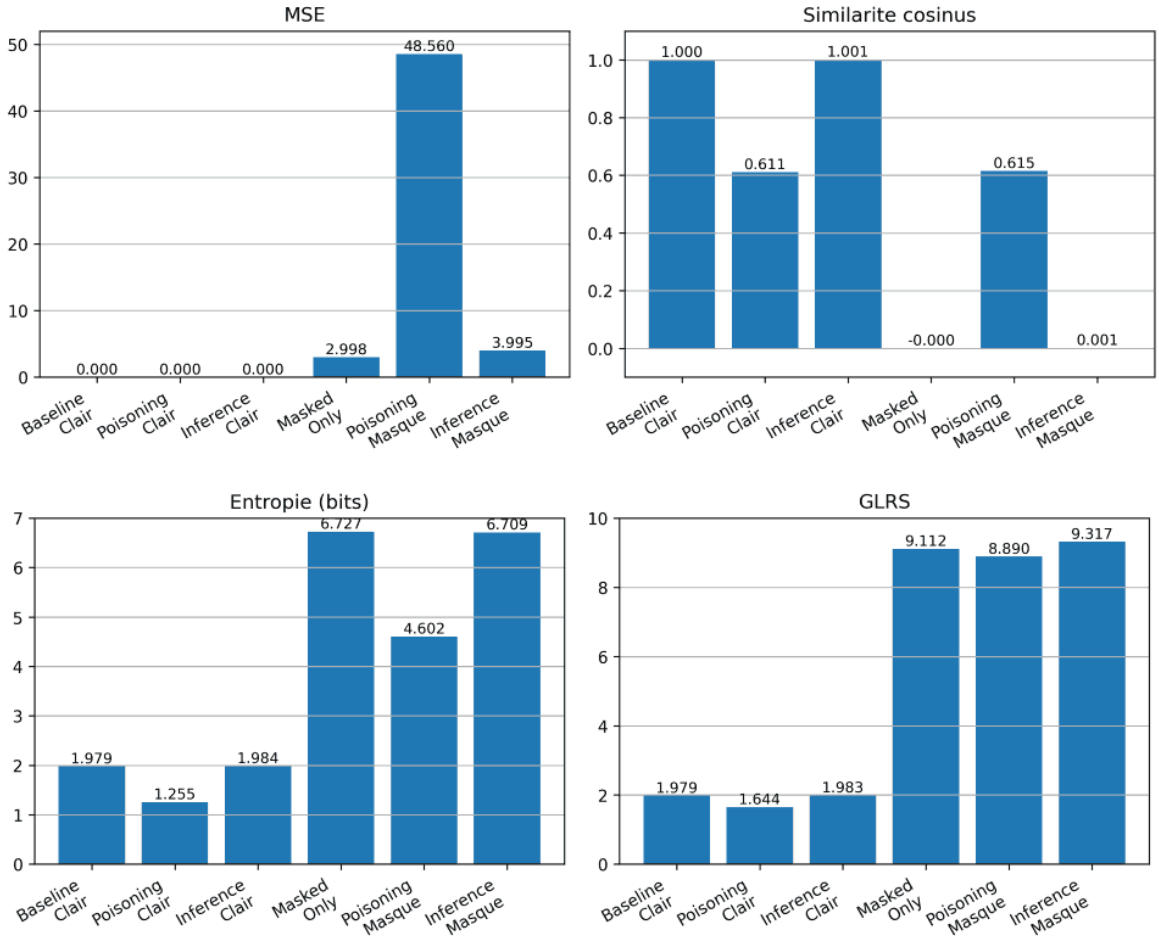

FIGURE 4.11 : Évaluation des mécanismes de sécurité de SSDLite

Sur les quatre sous figures 4.11 (MSE, similarité cosinus, entropie, GLRS), la baseline montre des gradients non protégés (MSE ≈ 0, cosinus ≈ 1, GLRS ≈ 1,98). Après masquage

côté client, le MSE grimpe à ≈ 48,560, la similarité cosinus tombe à ≈ 0 et l'entropie atteint ≈ 6,3–6,8 bits, ce qui indique des gradients très fortement distordus et quasi aléatoires, y compris sous poisoning et inférence. Le GLRS augmente jusqu'à ≈ 9,11–9,32, record parmi les modèles testés, ce qui reflète une excellente résilience aux fuites d'informations et rend les attaques de reconstruction ou de poisoning particulièrement difficiles à exploiter Wang & Bovik (2009); Zhu *et al.* (2019); Shannon (1948); Khatri (2012).

Les résultats démontrent que les gradients protégés diffèrent énormément des gradients bruts, ce qui rend leur utilisation par des attaques de reconstruction ou d'inférence (comme DLG/iDLG) beaucoup plus complexe. En utilisant à la fois le masquage pairwise et le bruit gaussien, on diminue considérablement l'information utile des gradients, tout en permettant au modèle global de converger correctement lors de l'agrégation fédérée. Cependant, cette évaluation demeure empirique et ne constitue pas une preuve formelle de sécurité cryptographique pour l'ensemble du mécanisme de sécurité Zuo *et al.* (2022); Bonawitz *et al.* (2017); Yang *et al.* (2022); Zhu *et al.* (2019).

4.6 IMPACT DES MÉCANISMES DE SÉCURITÉ

L'intégration de mécanismes de sécurité avancés dans un système d'apprentissage fédéré entraîne inévitablement un surcoût computationnel et communicationnel. Cette section permet d'analyser de manière quantitative l'impact des mécanismes de sécurité proposés (masquage pairwise, ajout de bruit gaussien, chiffrement symétrique et signatures numériques) sur les performances globales du système. L'objectif de ce mécanisme est d'évaluer le compromis entre le niveau de sécurité obtenu et l'efficacité opérationnelle, à partir de mesures expérimentales réelles Zhang *et al.* (2021).

4.6.1 COÛT COMPUTATIONNEL

Le coût computationnel causé par le mécanisme sécurisé est calculé en se basant sur les temps d'exécution mesurés lors d'un round fédéré complet, en particulier au round 10, correspondant à un état de convergence avancé du modèle. Les résultats démontrent que la charge de calcul est largement dominée par l'entraînement local du modèle de détection d'objets, tandis que les opérations cryptographiques n'introduisent qu'un surcoût marginal Bonawitz *et al.* (2017).

Ces résultats démontrent que les opérations de chiffrement, de signature et de vérification représentent une fraction négligeable du temps total d'exécution, comparée à l'entraînement du modèle profond. Le mécanisme sécurisé reste ainsi compatible avec des architectures exigeantes telles que YOLOv8, sans compromettre la faisabilité pratique du système dans un contexte réel d'apprentissage fédéré Li *et al.* (2024)

4.6.2 COÛT DE COMMUNICATION

Le chiffrement des gradients protégés et leur encodage en Base64 entraînent une augmentation de la taille des données échangées entre les différents clients et le serveur. Ce surcoût de communication est quantifié à partir des fichiers transmis lors des expérimentations. Bien que la taille des gradients chiffrés soit supérieure à celle des gradients non protégés, ce surcoût reste maîtrisé et acceptable dans des environnements réseau modernes. Il est largement compensé par le gain significatif en confidentialité et en robustesse face aux attaques par reconstruction et inférence Zuo *et al.* (2022)

4.7 CONCLUSION DU CHAPITRE

Ce chapitre a présenté une analyse expérimentale complète et approfondie de l'impact du mécanisme de sécurité post-quantique proposé sur la confidentialité des gradients, la robustesse face aux attaques et les performances globales du modèle de détection d'objets. Cette évaluation s'est appuyée sur un ensemble cohérent des métriques statistiques, informationnelles et de performance, permettant ainsi d'examiner simultanément les dimensions sécurité et efficacité du système. Les résultats expérimentaux nous démontrent que la combinaison des mécanismes de protection qui sont intégrés tels que (masquage pairwise, ajout de bruit gaussien, chiffrement symétrique et signatures numériques) permet de contrer efficacement les différentes attaques de reconstruction de gradients, d'inférence et de corrélation statistique. L'augmentation significative de l'entropie, la chute drastique de la similarité cosinus et des corrélations de Spearman et de Kendall, ainsi que l'élévation de la MSE confirment la une réduction quasi totale des fuites d'information exploitables à partir des gradients locaux. En outre, l'analyse des performances montre que ces garanties de confidentialité n'entraînent pas de dégradation notable de l'apprentissage fédéré. Le modèle YOLOv8 conserve une convergence stable et des performances de détection élevées, comme l'illustrent les courbes de précision, de rappel, de F1-score et la matrice de confusion normalisée. Ces observations confirment que les mécanismes de sécurité proposés sont compatibles avec des modèles de vision par ordinateur profonds et exigeants. L'étude des coûts computationnels et communicationnels met également en évidence que le surcoût induit par les opérations cryptographiques demeure marginal par rapport au temps d'entraînement local du modèle. Ce résultat est particulièrement important dans un contexte fédéré, car il confirme que l'intégration de mécanismes de sécurité avancés reste compatible avec un déploiement opérationnel du mécanisme proposé dans des environnements réels. En synthèse, les résultats présentés dans ce chapitre valident expérimentalement la faisabilité, l'efficacité et la pertinence du mécanisme de sécurisation de

l'apprentissage fédéré proposé. Ils montrent qu'il est possible de renforcer significativement la confidentialité des mises à jour locales et la résistance aux attaques, tout en maintenant des performances prédictives élevées et une efficacité opérationnelle acceptable. Ces conclusions constituent une base solide pour la discussion générale et l'analyse critique développées dans le chapitre suivant.

CHAPITRE V

DISCUSSION

Cette section discute les résultats expérimentaux obtenus au regard des objectifs de sécurité, de confidentialité et de performance définis dans la méthodologie, ainsi que leur positionnement par rapport à l'état de l'art en apprentissage fédéré sécurisé et post- quantique. Contrairement à la section VI, qui se concentre sur la présentation factuelle des mesures, cette discussion a pour but d'interpréter les résultats, d'analyser leurs implications et de mettre en perspective les contributions du mécanisme proposé.

5.1 SECURE AGRÉGATION ET GARANTIES DE CONFIDENTIALITÉ

Les résultats présentés dans la section VI montrent que le mécanisme de sécurité proposé empêche efficacement le serveur d'accéder aux gradients individuels en clair, conformément au modèle de menace du serveur honnête, mais curieux largement adopté en apprentissage fédéré sécurisé Zhang *et al.* (2021). Grâce au masquage pairwise, les contributions locales sont protégées de manière distribuée et les masques s'annulent mathématiquement lors de l'agrégation, garantissant que seule une mise à jour globale est observable côté serveur, comme dans les mécanismes d'agrégation sécurisée de référence Bonawitz *et al.* (2017).

En ajoutant du bruit gaussien aux gradients masqués, on renforce cette protection en réduisant de manière drastique les dépendances statistiques résiduelles entre gradients bruts et gradients partagés. Les valeurs très faibles de la similarité cosinus et des corrélations de Spearman et de Kendall témoignent d'une rupture quasi totale de la structure exploitable des gradients. Ces observations suggèrent une réduction significative de la surface d'attaque des méthodes de reconstruction de type Deep Leakage from Gradients (DLG) et iDLG, qui

reposent sur l'alignement directionnel et les corrélations statistiques fines entre gradients Zhu *et al.* (2019).

Contrairement aux approches se limitant au chiffrement du canal de communication, le mécanisme proposé agit directement sur le contenu informationnel des gradients, ce qui constitue une propriété essentielle dans un contexte où le serveur peut être curieux tout en le respectant Zuo *et al.* (2022); Bonawitz *et al.* (2017).

5.2 APPORT DE LA CRYPTOGRAPHIE POST-QUANTIQUE DANS L'APPRENTISSAGE FÉDÉRÉ

Ce travail met l'accent sur l'intégration explicite de mécanismes post-quantiques dans le pipeline d'apprentissage fédéré. En utilisant Kyber, il est possible d'assurer une confidentialité à long terme des échanges, même en présence d'adversaires dotés de capacités de calcul quantique, conformément aux recommandations du NIST Chen *et al.* (2016).

La dérivation de clés via HKDF et l'utilisation d'un chiffrement symétrique pour la transmission des mises à jour permettent d'obtenir un bon compromis entre sécurité cryptographique et efficacité pratique Frankel *et al.* (2003). Les résultats expérimentaux montrent que ces mécanismes n'introduisent qu'un surcoût marginal par rapport au coût dominant de l'entraînement local des modèles de détection d'objets, ce qui confirme la faisabilité d'un apprentissage fédéré sécurisé post-quantique dans des environnements réels Zuo *et al.* (2022); Li *et al.* (2024).

Dans ce travail, les schémas CRYSTALS-Kyber et CRYSTALS-Dilithium sont considérés comme sûrs au sens des définitions standardisées par le NIST, et leur sécurité n'est pas réévaluée via les métriques statistiques utilisées sur les gradients Bernstein & Lange (2017)

5.3 IMPACT DES MÉCANISMES DE SÉCURITÉ SUR LA CONVERGENCE ET LES PERFORMANCES

Les mécanismes de protection ont un impact potentiel sur la convergence et les performances du modèle global, ce qui constitue un enjeu majeur pour les systèmes d'apprentissage fédéré sécurisé. Les résultats obtenus montrent que la convergence du modèle fédéré reste stable et conforme aux attentes de l'algorithme FedAvg même en utilisant le masquage pairwise, le bruit gaussien et les opérations cryptographiques Zuo *et al.* (2022); Li *et al.* (2024).

Les performances de détection observées suggèrent que les mécanismes de sécurité ne perturbent pas de manière significative les capacités prédictives du modèle global. Ce constat est particulièrement important pour les architectures de détection d'objets profonds, tels que YOLO et SSD, connus pour leur sensibilité aux perturbations des gradients. Ces résultats confirment que l'intégration de la protection des gradients peut garantir l'efficacité du processus d'apprentissage fédéré Redmon *et al.* (2015); Sawatzky *et al.* (2019).

5.3.1 GÉNÉRALISATION À PLUSIEURS ARCHITECTURES DE DÉTECTION

Un autre point fort de cette étude réside dans l'évaluation du mécanisme sur plusieurs architectures de détection d'objets, à savoir le modèle YOLO principal, YOLO11n et SSD. Les résultats montrent que les tendances observées en ce qui concerne la confidentialité des gradients, la stabilité de convergence et les performances de détection sont similaires entre ces architectures.

Cette généralisation montre que la méthode suggérée n'est pas spécifique à une seule variante de YOLO, mais qu'elle s'applique efficacement à divers modèles de vision par ordinateur. Elle répond de manière explicite aux critiques concernant une évaluation restreinte à une seule architecture, et renforce la portée des conclusions expérimentales.

5.3.2 POSITIONNEMENT PAR RAPPORT À L'ÉTAT DE L'ART

Par rapport à l'état de l'art, le mécanisme proposé se distingue par la combinaison cohérente de plusieurs mécanismes complémentaires : L'agrégation sécurisée par masquage pairwise, protection statistique des gradients par ajout de bruit, et sécurisation des communications à l'aide de primitives post-quantiques. De nombreux travaux actuels se focalisent sur un seul de ces aspects, souvent au détriment des autres, ou ne garantissent que partiellement la confidentialité Zuo *et al.* (2022); Li *et al.* (2024); Bonawitz *et al.* (2017).

Les résultats expérimentaux montrent que cette approche multicouche permet d'obtenir un niveau élevé de confidentialité des gradients, sans compromettre la convergence ni les performances du modèle global. Ce choix entre sécurité élevée et réalisabilité concrète est un avantage pertinent pour des applications sensibles de l'apprentissage en réseau, notamment en vision par ordinateur distribuée Zhang *et al.* (2021).

5.3.3 LA COMPARAISON AVEC LA CONFIDENTIALITÉ DIFFÉRENTIELLE (DP) ET LE CHIFFREMENT HOMOMORPHE (HE)

Ce mécanisme d'agrégation sécurisée diffère fondamentalement des approches de confidentialité différentielle (DP) et de chiffrement homomorphe (HE) :

1. La Differential Privacy (DP), L'ajout un bruit aléatoire permanent aux gradients pour assurer la confidentialité. Ce bruit ne s'annule pas après l'agrégation, ce qui peut entraîner une dégradation de l'utilité du modèle global lorsque la confidentialité est forte (epsilon faible) Shannon (1948).
2. Le chiffrement homomorphe (HE), notamment basé sur le cryptosystème de Paillier, chiffre les gradients et permet au serveur d'effectuer l'agrégation sans jamais les déchif-

frer. Cette solution offre une confidentialité cryptographique forte, mais avec un coût computationnel très élevé et une forte latence, ce qui la rend difficilement utilisable dans un système d'apprentissage fédéré comportant de nombreux clients Paillier (1999).

Notre approche utilise un masque pairwise construit à partir d'un bruit gaussien déterministe, appliqué avec un signe opposé entre voisins, permettant son annulation mathématique après FedAvg. Ainsi, le serveur agrège les gradients en clair uniquement après l'annulation du masque, sans bruit permanent ni calcul cryptographique lourd en ligne Bonawitz *et al.* (2017).

Les résultats expérimentaux montrent que le mécanisme de protection des gradients entraîne une augmentation significative de l'entropie et une réduction drastique des métriques de similarité et de corrélation, notamment la similarité cosinus ainsi que les coefficients de Spearman et de Kendall, ce qui rend l'information exploitable par les attaques de fuite et de reconstruction de gradients, telles que Deep Leakage from Gradients (DLG) et iDLG, presque entièrement supprimée. Parallèlement, l'apprentissage fédéré conserve une convergence stable et les performances du modèle global ne présentent aucune dégradation notable, ce qui confirme l'efficacité pratique de l'agrégation sécurisée sans ajout de bruit permanent ni calcul cryptographique lourd en ligne. L'évaluation du mécanisme sur plusieurs architectures de détection d'objets, notamment YOLO (modèle principal), YOLOv11n et SSDLite montrent que les tendances de confidentialité des gradients et de performances de détection demeurent cohérentes entre les modèles, démontrant ainsi qu'il n'est pas spécifique à une seule architecture, mais généralisable et transposable à différents détecteurs profonds dans un cadre d'apprentissage fédéré sécurisé et résistant aux menaces quantiques Redmon *et al.* (2015); Zhu *et al.* (2019); Zhang *et al.* (2022).

Une formalisation complète du schéma, ainsi qu'une preuve de sécurité cryptographique dans un modèle standard constituent une perspective de travaux futurs.

CONCLUSION

Dans cette étude, nous avons présenté un mécanisme d'apprentissage fédéré sécurisé post-quantique. Nous avons évalué ce mécanisme avec un scénario de détection des objets en utilisant des modèles de vision profonde. Contrairement à de nombreux travaux qui se limitent à la sécurisation du canal de communication ou à la comparaison isolée d'algorithmes de cryptographie post-quantique, la contribution principale de cette étude réside dans la conception et la validation expérimentale d'un mécanisme d'agrégation sécurisée post-quantique, protégeant directement le contenu informationnel des gradients échangés en apprentissage fédéré.

Le mécanisme de sécurité proposé intègre de manière cohérente plusieurs composants complémentaires : (i) L'établissement des clés résistant au calcul quantique qui sont basé sur CRYSTALS-Kyber (ii) un chiffrement symétrique qui est dérivé via HKDF, (iii) L'authentification des échanges (iv) Le masquage pairwise des gradients, assurant ainsi leur annulation mathématique lors de l'agrégation FedAvg, et (v) l'ajout d'un bruit gaussien déterministe, utilisés comme masque temporaire annulable, afin de briser les corrélations statistiques qui sont exploitables avant l'agrégation. Cette approche nous garantit que le serveur, même honnête, n'a jamais accès aux gradients individuels en clair.

BIBLIOGRAPHIE

Abdi, H. (2007). The Kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*.

AbhishekV, A., Binny, S., JohanT, R., Raj, N. & Thomas, V. (2022). Federated Learning : Collaborative Machine Learning without Centralized Training Data. *international journal of engineering technology and management sciences*.

Arya, C., Tripathi, A., Singh, P., Diwakar, M., Sharma, K., Pandey, H. *et al.* (2021). Object detection using deep learning : A review. Dans *Journal of Physics : Conference Series*, Vol. 1854, p. 012012. IOP Publishing.

Bernstein, D. J. & Lange, T. (2017). Post-quantum cryptography—dealing with the fallout of physics success. *Cryptology ePrint Archive*.

Beullens, W., D'Anvers, J.-P., Hülsing, A. T., Lange, T., Panny, L., de Saint Guilhem, C. & Smart, N. P. (2021). Post-Quantum Cryptography : Current state and quantum mitigation.

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., de Gusmão, P. P. B. *et al.* (2020). Flower : A friendly federated learning research framework. *arXiv*.

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

Chen, L., Chen, L., Jordan, S., Liu, Y.-K., Moody, D., Peralta, R., Perlner, R. A. & Smith-Tone, D. (2016). *Report on post-quantum cryptography*. US Department of Commerce, National Institute of Standards and Technology.

Chen, L., Moody, D. & Liu, Y. (2017). NIST post-quantum cryptography standardization. *Transition*.

Das, S. & Das, A. (2025). Pre-Quantum to Post-Quantum Cryptography Transition : A

Journey Connecting the Security and Challenges Eras. Dans *Integration of AI, Quantum Computing, and Semiconductor Technology*. IGI Global.

Dekkaki, K. C., Tasic, I. & Cano, M.-D. (2024). Exploring post-quantum cryptography : Review and directions for the transition process. *Technologies*.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K. & Leal-Taixé, L. (2020). Mot20 : A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv :2003.09003*.

Frankel, S., Glenn, R. & Kelly, S. (2003). *The AES-CBC Cipher Algorithm and Its Use with IPsec* publication n° NIST Special Publication 800-38A. National Institute of Standards and Technology.

Gorbenko, I. & Kandii, S. (2025). National and International Post-Quantum Standards for Asymmetric Transformations. *Cybernetics and Systems Analysis*, *61*(4), 659–670.

Gurung, D., Pokhrel, S. R. & Li, G. (2023). Secure communication model for quantum federated learning : A post quantum cryptography (pqc) framework. *arXiv*.

Hauke, J. & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*.

Heik, D., Bahrpeyma, F. & Reichelt, D. (2025). YoloRL : simplifying dynamic scheduling through efficient action selection based on multi-agent reinforcement learning.

Joseph, D., Misoczki, R., Manzano, M., Tricot, J., Pinuaga, F. D., Lacombe, O., Leichenauer, S., Hidary, J., Venables, P. & Hansen, R. (2022). Transitioning organizations to post-quantum cryptography. *Nature*, *605*(7909), 237–243.

Kaur, J. & Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image : a review. *Multimedia Tools and Applications*.

Khatri, M. (2012). Cosine similarity function for the temporal dynamic web data. *International Journal of Computer Science & Engineering Technology*.

Li, P., Chen, T. & Liu, J. (2024). Enhancing Quantum Security over Federated Learning via Post-Quantum Cryptography. Dans *IEEE TPS-ISA*.

Liu, B., Lv, N., Guo, Y. & Li, Y. (2024). Recent advances on federated learning : A systematic survey. *Neurocomputing*, p. 128019.

McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Dans *Artificial intelligence and statistics*.

Moore, B. E. & Corso, J. J. (2020). FiftyOne. *GitHub. Note : https ://github.com/voxel51/fiftyone*.

Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. Dans *International conference on the theory and applications of cryptographic techniques*.

Redmon, J., Divvala, S. K., Girshick, R. B. & Farhadi, A. (2015). You Only Look Once : Unified, Real-Time Object Detection. *IEEE CVPR*.

Salowey, J., Choudhury, A. & McGrew, D. (2008). *AES Galois Counter Mode (GCM) Cipher Suites for TLS* publication n° RFC 5288. Internet Engineering Task Force.

Sawatzky, J., Souri, Y., Grund, C. & Gall, J. (2019). What Object Should I Use ? - Task Driven Object Detection. *Conference on Computer Vision and Pattern Recognition*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*.

Shor, P. W. (1999). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*.

Tasnim, S. & Qi, W. (2023). Progress in Object Detection : An In-Depth Analysis of Methods and Use Cases. *European Journal of Electrical Engineering and Computer Science*.

Ultralytics (2024). *YOLO Performance Metrics*. Accessed : 2025-02-05, Repéré à https://docs.ultralytics.com/guides/yolo-performance-metrics/

Wang, X., Gao, H., Jia, Z. & Li, Z. (2023). BL-YOLOv8 : An improved road defect detection model based on YOLOv8. *Sensors*, *23*(20), 8361.

Wang, Z. & Bovik, A. C. (2009). Mean squared error : Love it or leave it ? A new look at signal fidelity measures. *IEEE signal processing magazine*.

Yang, S., Chen, Y., Tu, S. & Yang, Z. (2022). A post-quantum secure aggregation for federated learning. Dans *12th International Conference on Communication and Network Security*.

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D. & Beaufays, F. (2018). Applied federated learning : Improving google keyboard query suggestions. arXiv 2018. *arXiv preprint arXiv :1812.02903*.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W. & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*.

Zhang, J., Jing, J., Lu, P. & Song, S. (2022). Improved MobileNetV2-SSDLite for automatic fabric defect detection system based on cloud-edge computing. *Measurement*, *201*, 111665.

Zhu, L., Liu, Z. & Han, S. (2019). Deep leakage from gradients. *Advances in neural information processing systems*, *32*.

Zuo, R., Tian, H., An, Z. & Zhang, F. (2022). Post-quantum privacy-preserving aggregation in federated learning based on lattice. Dans *International Symposium on Cyberspace Safety and Security*.