الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbes

Graduation Thesis

To obtain the diploma of **Engineering Degree**

Field of Study: **Computer Science**

Specialization: **artificial intelligence and data science**

Theme

Privacy-Preserving Federated Random Forest for Healthcare Applications

---

Presented by
Mohamed El Amine Serradj

Defended on: **09,2025**
In front of the jury composed of

Mr. BENDELLA Mohammed Salih	President of the Jury
Mr. Miloud Khaldi	Thesis Supervisor
Mr. ELHANNANI Souad	Examiner

Academic Year: 2024/2025

Abstract

The rapid growth of data-driven artificial intelligence in sensitive domains such as healthcare has raised urgent concerns about privacy, security, and regulatory compliance. Traditional centralized learning is increasingly infeasible under legal frameworks such as GDPR, motivating the adoption of *Federated Learning (FL)*, which enables multiple institutions to collaboratively train models without sharing raw data.

Among machine learning methods, tree ensembles—including Random Forests (RF) and Gradient Boosted Decision Trees (GBDTs)—are especially relevant due to their interpretability, robustness, and predictive performance. However, training such models in federated settings poses unique challenges, as split selection and leaf evaluation require data-dependent computations that risk privacy leakage if performed naively.

Two major families of privacy-preserving techniques have dominated recent research. **Differential Privacy (DP)** introduces calibrated randomness to protect against inference attacks, providing quantifiable guarantees at the record level. **Secure Multi-Party Computation (MPC)** enables multiple parties to compute aggregate statistics securely without exposing local data. While both approaches have shown promise independently, their integration remains underexplored, particularly for Random Forests.

The state of the art (2020–2025) reveals complementary baselines: MPC-only methods such as FederBoost and FedTree achieve high utility but lack formal output privacy, while DP-only methods such as those of Maddock, Marković, and Tao provide record-level guarantees but often suffer from reduced accuracy and incomplete privacy accounting. No current approach simultaneously achieves DP and MPC in fully decentralized federated settings, leaving important gaps in scalability, noise placement strategies, and rigorous privacy composition.

This thesis reviews these advances, identifying open challenges and outlining future research directions. Key opportunities include hybrid DP+MPC protocols, decentralized orchestration without central servers, systematic analysis of noise placement, and reproducibility through open-source implementations. Addressing these challenges is essential to enable trustworthy, privacy-preserving federated learning for real-world applications in healthcare and beyond.

Keywords— Federated Learning, Privacy-Preserving Machine Learning,Random Forest,Differential Privacy,Secure Multi-Party Computation,Decentralized Systems

Résumé

La croissance rapide de l'intelligence artificielle fondée sur les données dans des domaines sensibles tels que la santé a soulevé des préoccupations majeures en matière de confidentialité, de sécurité et de conformité réglementaire. L'apprentissage centralisé traditionnel devient de plus en plus difficile à mettre en œuvre dans le cadre juridique imposé par des réglementations comme le RGPD, ce qui motive l'adoption de l'*Apprentissage Fédéré (Federated Learning, FL)*, permettant à plusieurs institutions de collaborer pour entraîner des modèles sans partager directement leurs données brutes.

Parmi les méthodes d'apprentissage, les ensembles d'arbres—incluant les Forêts Aléatoires (Random Forests, RF) et les Arbres de Décision Boostés (Gradient Boosted Decision Trees, GBDTs)—occupent une place centrale grâce à leur interprétabilité, leur robustesse et leurs performances prédictives. Cependant, l'entraînement de ces modèles en contexte fédéré pose des défis spécifiques, car la sélection des divisions et l'évaluation des feuilles impliquent des calculs dépendants des données susceptibles de compromettre la confidentialité s'ils sont réalisés naïvement.

Deux grandes familles de techniques de préservation de la vie privée dominent la recherche actuelle. **La Confidentialité Différentielle (Differential Privacy, DP)** ajoute un bruit calibré afin de limiter les attaques d'inférence, garantissant la protection au niveau des enregistrements individuels. **Le Calcul Sécurisé Multi-Parties (Secure Multi-Party Computation, MPC)** permet à plusieurs entités de calculer des statistiques globales de manière sécurisée sans exposer leurs données locales. Bien que ces deux approches aient montré leur efficacité séparément, leur intégration reste encore peu étudiée, en particulier pour les forêts aléatoires.

L'état de l'art (2020–2025) révèle des lignes de base complémentaires : les méthodes fondées uniquement sur le MPC, comme FederBoost et FedTree, atteignent une utilité élevée mais sans garanties de confidentialité formelles sur les sorties, tandis que les méthodes basées uniquement sur la DP, telles que celles de Maddock, Marković et Tao, offrent des garanties au niveau des enregistrements mais subissent souvent une baisse d'exactitude et un manque d'analyse complète de la composition de la confidentialité. Aucun travail existant n'atteint aujourd'hui une combinaison DP+MPC dans un cadre fédéré pleinement décentralisé, ce qui laisse des lacunes importantes en matière de passage à l'échelle, de stratégies d'injection de bruit et de rigueur dans l'analyse de la confidentialité.

Ce mémoire passe en revue ces avancées, identifie les défis ouverts et trace des perspectives de recherche futures. Parmi celles-ci figurent le développement de protocoles hybrides DP+MPC, l'orchestration décentralisée sans serveur central, l'étude systématique du placement du bruit, ainsi que la reproductibilité grâce à des implémentations open-source. Relever ces défis est essentiel pour permettre un apprentissage fédéré respectueux de la vie privée, digne de confiance et applicable aux domaines sensibles comme la santé et la finance.

Keywords— Apprentissage Fédéré, Forêt Aléatoire,Confidentialité Différentielle,Calcul Multi-Party Sécurisé,Systèmes Décentralisés