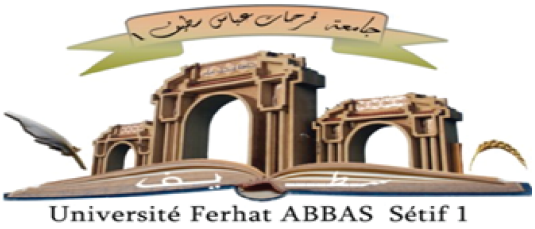
**People's Democratic Republic of Algeria**

**Ministry of Higher Education and Scientific Research**

**University Sétif 1 – Ferhat Abbas**

**Faculty of Science**

**Department of Computer Science**




# Master's Degree Thesis

## in Computer Science

***Specialization:***

*Cybersecurity*

*Data Engineering and Web Technology*

# Topic

---

## Detection and Mitigation of Adversarial Attacks in Federated Learning for IoT Networks

---

**Supervised by:**

Pr. ALIOUAT née ZOUAOUI Zibouda

KHACHA Amina

**Prepared by:**

KADRI Haroune

CHELLALI Baha Eddine

2025/2026

# Contents

## Abstract

The rapid expansion of the Internet of Things (IoT) has transformed data collection across critical sectors, particularly healthcare. While Federated Learning (FL) offers a promising paradigm by enabling collaborative model training without centralizing sensitive data, its distributed architecture introduces significant security and privacy vulnerabilities. This thesis investigates the detection and mitigation of adversarial attacks in FL systems deployed over IoT networks. We focus on a realistic cross-silo medical scenario in which multiple hospitals collaboratively train a chest X-ray disease detection model based on the DenseNet-121 architecture. We first demonstrate the vulnerability of standard FL to white-box Membership Inference Attacks (MIA), revealing how data overlap between training and test sets creates exploitable structural backdoors that enable an attacker to determine whether a specific patient's data was used during training. To counter these threats, we propose a progressive defense-in-depth strategy combining three complementary layers: (1) structural data reorganization through patient-level splitting combined with Low-Rank Adaptation (LoRA) fine-tuning, (2) Selective Differential Privacy with temperature scaling for mathematical noise injection, and (3) Homomorphic Encryption (HE) for securing the communication channel between clients and the server. Experimental results demonstrate that our integrated layered defense effectively reduces the attacker's success rate to random chance while preserving the diagnostic performance of the model, proving that secure and privacy-preserving collaborative AI over IoT networks is practically achievable.

**Keywords:** Federated Learning, Internet of Things, Adversarial Attacks, Membership Inference Attack, Differential Privacy, Homomorphic Encryption, Medical Image Classification, Deep Learning.

## Résumé

L'expansion rapide de l'Internet des Objets (IoT) a transformé la collecte de données dans les secteurs critiques, en particulier la santé. Bien que l'Apprentissage Fédéré (FL) offre un paradigme prometteur en permettant l'entraînement collaboratif de modèles sans centraliser les données sensibles, son architecture distribuée introduit des vulnérabilités

significatives en matière de sécurité et de confidentialité. Ce mémoire étudie la détection et l'atténuation des attaques adversariales dans les systèmes d'apprentissage fédéré déployés sur les réseaux IoT. Nous nous concentrons sur un scénario médical réaliste de type cross-silo, dans lequel plusieurs hôpitaux entraînent collaborativement un modèle de détection de maladies à partir de radiographies thoraciques, basé sur l'architecture DenseNet-121. Nous démontrons d'abord la vulnérabilité du FL standard aux attaques par inférence d'appartenance (MIA) en boîte blanche, révélant comment le chevauchement des données entre les ensembles d'entraînement et de test crée des portes dérobées structurelles exploitables permettant à un attaquant de déterminer si les données d'un patient spécifique ont été utilisées lors de l'entraînement. Pour contrer ces menaces, nous proposons une stratégie de défense en profondeur progressive combinant trois couches complémentaires : (1) la réorganisation structurelle des données par séparation au niveau patient combinée à l'adaptation de faible rang (LoRA), (2) la Confidentialité Différentielle Sélective avec mise à l'échelle de température pour l'injection de bruit mathématique, et (3) le Chiffrement Homomorphe (HE) pour sécuriser le canal de communication entre les clients et le serveur. Les résultats expérimentaux démontrent que notre défense multicouche intégrée réduit efficacement le taux de succès de l'attaquant au niveau du hasard tout en préservant les performances diagnostiques du modèle, prouvant que l'IA collaborative sécurisée et respectueuse de la vie privée sur les réseaux IoT est pratiquement réalisable.

**Mots clés :** Apprentissage Fédéré, Internet des Objets, Attaques Adversariales, Attaque par Inférence d'Appartenance, Confidentialité Différentielle, Chiffrement Homomorphe, Classification d'Images Médicales, Apprentissage Profond.

## ملخص

أدّى التوسع السريع لإنترنت الأشياء إلى تحويل عملية جمع البيانات في القطاعات الحيوية، وخاصةً الرعاية الصحية. رغم أن التعلم الموحّد يُقدّم نموذجاً واعداً من خلال تمكين التدريب التعاوني للنماذج دون مركزة البيانات الحساسة، إلا أن بنيته الموزعة تُدخل ثغرات أمنية كبيرة تتعلق بالأمان والخصوصية. تبحث هذه المذكرة في كشف والتخفيف من الهجمات العدائية في أنظمة التعلم الموحّد المنشورة على شبكات إنترنت الأشياء. نركّز على سيناريو طبي واقعي من نوع cross-silo حيث تتعاون عدة مستشفيات لتدريب نموذج للكشف عن الأمراض من صور الأشعة السينية للصدر باستخدام بنية DenseNet-121 . نُثبت أولاً ضعف التعلم الموحّد القياسي أمام هجمات استنتاج العضوية من نوع الصندوق الأبيض، كاشفين كيف يُنشئ تداخل البيانات بين مجموعات التدريب والاختبار أبواباً خلفية هيكلية قابلة للاستغلال تُمكّن المهاجم من تحديد ما إذا كانت بيانات مريض معيّن قد استُخدمت أثناء التدريب. لمواجهة هذه التهديدات، نقترح استراتيجية دفاع متعددة الطبقات تجمع بين ثلاث طبقات متكاملة: أولاً إعادة التنظيم الهيكلي للبيانات عبر الفصل على مستوى المريض مع التكيّف منخفض الرتبة، ثانياً الخصوصية التفاضلية الانتقائية مع تعديل درجة الحرارة لحقن الضوضاء الرياضية، وثالثاً التشفير المتجانس لتأمين قناة الاتصال بين العملاء والخادم. تُظهر النتائج التجريبية أن دفاعنا المتكامل متعدد الطبقات يُقلّل بفعالية معدل نجاح المهاجم إلى مستوى الصدفة مع الحفاظ على الأداء التشخيصي للنموذج، مما يُثبت أن الذكاء الاصطناعي التعاوني الآمن والمحافظ على الخصوصية عبر شبكات إنترنت الأشياء قابل للتحقيق عملياً.



**الكلمات المفتاحية:** التعلم الموحّد، إنترنت الأشياء، الهجمات العدائية، هجوم استنتاج العضوية، الخصوصية التفاضلية، التشفير المتجانس، تصنيف الصور الطبية، التعلم العميق.