TITRE DE LA THESE:

Défenses et attaques par porte dérobée en apprentissage fédéré en santé /
Defenses And BackDoor attacks in federated healthcare learning

ACRONYME DE LA THESE:

DABaDoo

Direction de thèse :

Directeur de thèse: Gouenou Coatrieux, Professeur, Responsable équipe CYBER HEALTH, LaTIM, IMT Atlantique

Co-encadrant·es :

M. Kassem Kallas, Chercheur Inserm, équipe CYBER HEALTH, LaTIM, Spécialiste en sécurité de l'IA, Porteur au d'une action de recherche sur la lutte contre les attaques adversaires (e.g.les backdoor).

Laboratoire(s) :

GEPEA	IRISA	Lab-STICC	**X**	LATIM
Lego	LEMNA	LS2N		hors Laboratoire

Equipe(s) de recherche :

Équipe CYBER HEALTH au sein de LaTIM

Département(s) IMT Atlantique :

DAPI	DSEE	INFO	**X**	ITI	LCI	LUSSI
MEE	MO	OPT		SSG	SRCD	SUBATECH

S'agit-il d'une thèse en cotutelle internationale ?

Oui ☐ **Non ☑**

Si oui, organisme avec lequel la cotutelle est envisagée :

Le sujet proposé présente-il un caractère interdisciplinaire ?

Oui ☑ **Non ☐**

Si oui, expliquer brièvement pourquoi (2 ou 3 lignes) :

Le sujet proposé est interdisciplinaire car il combine les sciences de l'information et les sciences de la santé. Il explore des enjeux de sécurité dans l'apprentissage fédéré en santé. Il implique une collaboration étroite entre des experts en intelligence artificielle, en sécurité informatique et des professionnels de santé pour développer des solutions d'IA de confiance, robustes et fiables.

La source du co-financement est-elle identifiée ?

Oui ☑ **Non ☐**

Si oui, préciser quel co-financement est envisagé :

La source du co-financement est identifiée et provient du projet SSF-ML-DH (Secure, safe and fair machine learning for healthcare) du PEPR Santé Numérique du PIA Plan France 2030.

Autres informations :
Informations utiles que vous souhaiteriez communiquer (si pertinent) :
Cette demande a pour but de soutenir : i) un jeune chercheur Inserm qui a rejoint l'équipe LaTIM/Cyber Health el 01/12/2023, et lui permettre de développer rapidement son activité ; ii) une toute jeune équipe de recherche créée le 01/01/2022.

Contexte ou état de l'art scientifique :
Décrire en 5 à 10 lignes le contexte de la thèse.
Le domaine de l'apprentissage fédéré (« Federated Learning » - FL) en santé offre la promesse de développer des modèles d'intelligence artificielle (IA) robustes tout en respectant la confidentialité des données des patients. Cependant, cette stratégie rend les modèles vulnérables aux attaques de type porte dérobée (« backdoor ») [1] qui, lorsqu'elles sont activées, altèrent leur fonctionnement. Ce risque est inacceptable en santé avec des conséquences sur l'intégrité des décisions (dépistage, diagnostics, choix thérapeutique). La recherche actuelle sur ce type d'attaques et les défenses possibles dans un contexte fédéré est encore naissante. De nombreux défis sont à relever notamment en matière de détection et de neutralisation de ces attaques, comme le soulignent les études [2-4]. Les mécanismes de défense sont rares et peu testés dans le contexte médical, nécessitant une exploration approfondie et le développement de nouvelles stratégies défensives [5-7]. Cette thèse vise donc à approfondir la compréhension de ces menaces et à développer des mécanismes de défense efficaces, contribuant à l'avancement de l'IA de confiance en santé.

[1] Barni M, Kallas K, Tondi B. A new backdoor attack in cnns by training set corruption without label poisoning. In2019 IEEE International Conference on Image Processing (ICIP) 2019 Sep 22 (pp. 101-105). IEEE.
[2] J. H. Yoo, H. Jeong, J. Lee, and T. M. Chung, "Federated learning: Issues in medical application," in Future Data and Security Engineering: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings, vol. 8, Springer International Publishing, 2021, pp. 3-22.
[3] B. Xi, S. Li, J. Li, H. Liu, and H. Zhu, "Batfl: Backdoor detection on federated learning in e-health," in 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), IEEE, 2021, pp. 1-10.
[4] T. D. Nguyen, P. Rieger, R. De Viti, H. Chen, B. Brandenburg, H. Yalame, and T. Schneider, "{FLAME}: Taming backdoors in federated learning," in 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1415-1432.
[5] R. Bellafqira, G. Coatrieux, M. Lansari, J. Chala, FedCAM - Identifying Malicious Models in Federated Learning Environments Conditionally to Their Activation Maps, IEEE 19th Wireless On-Demand Network Systems and Services Conference (WONS),2024, pp.49-56.
[6] Y. Li, Y. Jiang, Z. Li, and S. T. Xia, "Backdoor learning: A survey," IEEE Transactions on Neural Networks and Learning Systems, 2022.
[7] Q. Le Roux, E. Bourbao, Y. Teglia, and K. Kassem, "A Comprehensive Survey on Backdoor Attacks and their Defenses in Face Recognition Systems," IEEE Access, vol. 12, pp. 47433-47468, 2024, doi: 10.1109/ACCESS.2024.3382584.

Objectifs de la thèse :
Décrire en 10 à 15 lignes les résultats attendus.
Les objectifs de cette thèse sont de deux ordres :
- Mener une analyse approfondie (état de l'art) des attaques de type backdoor en centralisé ou fédéré ; avec ou sans serveur d'agrégation des modèles ; et des défenses existantes. Cette analyse sera accompagnée de l'élaboration d'un banc de tests « attaque/défense » dans le contexte d'applications en santé. Il s'agit d'acquérir une compréhension approfondie de ces attaques et des défenses, et de favoriser la science ouverte, la reproductibilité et la comparaison des résultats.
- Concevoir de nouveaux mécanismes/stratégies de défense innovants et efficaces, intégrés à l'apprentissage fédéré ; i.e., des solutions évolutives permettant de détecter et bloquer ces attaques. Il s'agira notamment d'aller vers l'interprétabilité et la transparence des modèles de l'apprentissage fédéré, afin d'identifier les caractéristiques des modèles impactées par ces attaques et proposer des défenses robustes et résilientes. En plus de ces solutions fondées sur l'observation des données partagées entre participants, nous explorerons l'exploitation de techniques de tatouage de modèles.

Sont attendus des publications/communications dans des revues/conférences internationales majeures, contribuant à enrichir l'existant avec de nouvelles méthodes, des études de cas réelles et

des benchmarks qui serviront de référence pour les communautés scientifiques en cybersécurité et en IA en santé, sur un sujet très ouvert. La question de la valorisation industrielle sera aussi réfléchie.

Compétences attendues du ou de la candidat·e :
Lister les principales compétences nécessaires pour ce sujet de thèse.

- Solide formation académique dans un des domaines de la thèse (par exemple, informatique).
- Expérience(s) en apprentissage fédéré, en apprentissage automatique (Machine Learning) préservant la vie privée ou en cybersécurité est préférable.
- Compétence en programmation (Python, PyTorch, C++,) et en analyse de données.
- Une bonne compréhension des principes et des méthodologies de l'apprentissage fédéré sera très appréciée.
- Bonnes capacités d'analyse et de résolution de problèmes avec un souci du détail,
- Bonnes capacités de communication et de travail en équipe,
- Autonomie et esprit d’initiative,
- Une expérience dans le domaine de la santé (cardiologie ou neurosciences) serait un plus.