AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **M'HAOUACH Mohamed**
Soutiendra : **le Samedi 25/10/2025** à **14H30**
Lieu : **FSDM – *Centre Visioconférence***

Une thèse intitulée :

«Towards Intelligent Client Selection in Federated Learning for Robust and Scalable Artificial Intelligence Systems »

En vue d'obtenir le ***Doctorat***

FD : ***Sciences et Technologies de l'Information et de la Communication***
Spécialité : ***Intelligence Artificielle et Big Data***

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
ZINEDINE Ahmed	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Président
AL ACHHAB Mohammed	Ecole Nationale des Sciences Appliquées, Tétouan	PES	Rapporteur
EL KAMILI Mohamed	Ecole Supérieure de Technologie, Casablanca	PES	Rapporteur
BOUAYAD Anas	Faculté des Sciences Dhar EL Mahraz, Fès	MCH	Rapporteur
EL YADARI Mourad	École Nationale supérieure d'Arts et Métiers, Rabat	PES	Examinateur
BOUHOUTE Afaf	Faculté des Sciences Dhar EL Mahraz, Fès	MCH	Examinateur
FARDOUSSE Khalid	Faculté Chariaa, Fès	MCH	Directeur de thèse
BERRADA Ismail	Université Mohammed VI Polytechnique, Ben Guerir	MCH	Co-directeur de thèse

Résumé :

L'apprentissage fédéré (FL) permet à plusieurs dispositifs périphériques de collaborer sur un modèle partagé sans jamais transférer de données brutes, ce qui en fait une solution privilégiée pour les secteurs sensibles à la confidentialité tels que la santé, les télécommunications et les systèmes autonomes. En confinant les données aux smartphones, capteurs IoT et autres nœuds locaux, le FL atténue les risques pour la vie privée, réduit l'exposition aux cyberattaques et diminue la surcharge de bande passante typique des pipelines centralisés. Cette architecture décentralisée exploite également la diversité des jeux de données réels pour produire des modèles qui se généralisent mieux dans des environnements hétérogènes et contraints en latence.

Malgré ces avantages, le FL se heurte à d'importants obstacles lorsque les appareils clients détiennent des données non indépendantes et non identiquement distribuées (N-IID). Les stratégies standard de sélection des clients ne prennent pas en compte les variations de distribution des données, la puissance de calcul, la fiabilité du réseau et les contraintes énergétiques. Par conséquent, les cycles d'entraînement convergent lentement, les modèles agrégés héritent de biais statistiques et certaines sous-populations de données essentielles sont sous-représentées, ce qui dégrade la qualité globale du modèle dans des contextes fortement hétérogènes

Cette thèse doctorale aborde ces problématiques au travers de trois contributions :

1. **PyFed Framework.** Construit au-dessus de PySyft, PyFed est un environnement de simulation modulaire et extensible conçu pour un benchmarking rigoureux dans des conditions N-IID.
2. **NIFL (N-IID Index-based Federated Learning).** NIFL introduit un nouvel indice statistique quantifiant la similarité entre les distributions de données des clients sans exposer d'échantillons bruts. Cet indice N-IID oriente le serveur vers les clients dont les distributions sont les plus proches, réduisant ainsi les biais d'agrégation et accélérant la convergence tout en diminuant les coûts de communication.
3. **FREQSEL (Frequency-based Client Selection).** FREQSEL affine encore la sélection en se concentrant sur la fréquence des échantillons par classe. Les clients transmettent d'abord le nombre d'exemples par classe ; le serveur calcule ensuite la distance entre le vecteur de fréquences de chaque client et la distribution globale, ne conservant que ceux dont la divergence est minimale. Cette stratégie limite la dérive des poids lors de l'agrégation, offrant une convergence plus rapide, une meilleure précision et une efficacité accrue des ressources dans des réseaux hétérogènes.

Ensemble, PyFed, NIFL et FREQSEL font progresser l'état de l'art en apprentissage fédéré en fournissant : (i) une plateforme d'évaluation robuste et équitable, (ii) une approche statistique respectueuse de la confidentialité pour analyser l'hétérogénéité des données, et (iii) un mécanisme léger, conscient des distributions, pour la sélection des clients.

Mots clés : ***Apprentissage fédéré, Intelligence artificielle, Stratégies de sélection des clients, Apprentissage distribué, Informatique en edge mobile, Préservation de la confidentialité, Données non indépendantes et non identiquement distribuées (N-IID), Big data.***

TOWARDS INTELLIGENT CLIENT SELECTION IN FEDERATED LEARNING FOR ROBUST AND SCALABLE ARTIFICIAL INTELLIGENCE SYSTEMS

Abstract : Federated Learning (FL) enables multiple edge devices to collaborate on a shared model without ever transferring raw data, making it an attractive solution for privacy-sensitive sectors such as healthcare, telecommunications, and autonomous systems. By confining data to smartphones, IoT sensors, and other local nodes, FL mitigates privacy risks, reduces exposure to cyberattacks, and curbs the bandwidth overhead typical of centralized pipelines. The decentralized architecture also leverages the diversity of real-world datasets to yield models that generalize better in heterogeneous, latency-constrained environments.

Despite these advantages, FL faces severe obstacles when client devices hold data that is non-independent and non-identically distributed (N-IID). Standard client-selection strategies fail to account for variations in data distribution, computational power, network reliability, and energy reserves. As a result, training rounds converge slowly, aggregated models inherit statistical biases, and important data sub-populations are under-represented, eroding overall model quality in highly heterogeneous settings.

This doctoral dissertation tackles those issues through three contributions:

1. **PyFed Framework**. Built atop PySyft, PyFed is a modular and extensible simulation environment designed for rigorous benchmarking under N-IID conditions.
2. **NIFL (N-IID Index-based Federated Learning)**. NIFL introduces a new statistical index that quantifies the similarity among client data distributions without exposing raw samples. The N-IID index guides the server in choosing clients whose data distributions are closest, reducing aggregation bias and accelerating convergence while lowering communication overhead.
3. **FREQSEL (Frequency-based Client Selection)**. FREQSEL further refines selection by focusing on per-class sample frequencies. Clients first report class counts. The server then computes the distance between each client's class-frequency vector and the global distribution, retaining those with minimal divergence. This strategy limits weight drift during aggregation, yielding faster convergence, higher accuracy, and improved resource efficiency across heterogeneous networks.

Collectively, PyFed, NIFL, and FREQSEL advance the state of federated learning by providing (i) a robust testbed for fair comparison, (ii) a privacy-preserving statistical lens on data heterogeneity, and (iii) a lightweight, distribution-aware mechanism for client participation. Together, they offer a practical pathway toward scalable, accurate, and privacy-compliant FL in real-world, data-diverse ecosystems.

Key Words: ***Federated Learning, Artificial Intelligence, Clients Section Strategies, Distributed Learning, Mobile Edge Computing, Privacy-Preserving, Non-Independent and Identical Distributed (N-IID) Data, Big Data.***