**HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY**

# GRADUATION THESIS

## Federated Impurity Weighting: A Novel Approach for Improving Convergence in Federated Learning

**TẠ VIỆT CƯỜNG**
cuong.tv194422@sis.hust.edu.vn

**Major: Data Science and Artificial Intelligence**
**Specialization: Data Science and Artificial Intelligence**

**Supervisor:** Dr. Trần Hải Anh _____

Signature

**Department:** Computer Engineering

**School:** School of Information and Communications Technology

**HANOI, 01/2024**

# ACKNOWLEDGMENT

I would like to express my sincere gratitude to my two instructors, Dr. Hai-Anh Tran and Dr. Truong X. Tran, whose guidance and support have been instrumental throughout the journey of this research. Their wealth of knowledge, insightful feedback, and encouragement have been pivotal in shaping the course of my work, and I am truly fortunate to have had the opportunity to learn under their mentorship.

I also extend my heartfelt thanks to my family, especially my parents and sister, for their unwavering support, understanding, and encouragement. Their constant belief in my abilities and their willingness to stand by me during challenging times have been a tremendous source of motivation. Their love and encouragement have played a pivotal role in my academic pursuits.

Additionally, I want to acknowledge my friends whose camaraderie and encouragement have been a source of inspiration. Their shared enthusiasm for learning has made this journey even more rewarding, and I am grateful for the meaningful connections formed during this research endeavor.

# ABSTRACT

Federated Learning (FL) presents a collaborative learning approach for multiple devices, allowing them to jointly learn a global model while maintaining the privacy of locally stored data. However, challenges arise due to the non-Independently and Identically Distributed (non-IID) nature of data samples across nodes, leading to inefficiencies in model training and requiring additional communication rounds for convergence. Addressing the identified limitations of FedAdp in the context of non-IID data, this thesis introduces a novel FL algorithm named FedImp. Our investigation underscores the diverse contributions made by different nodes during the global model aggregation process. Building on this insight, our core approach involves measuring a participating node's contribution by evaluating the informational content within its data. These contributions are then normalized to generate distinct weights for the aggregation of the global model. Through systematic experimentation on diverse datasets such as EMNIST, CIFAR-10, and the Large Movie Review Dataset, FedImp consistently exhibits superior convergence speed. It achieves a significant reduction in the number of communication rounds required for convergence compared to both FedAdp and FedAvg. Empirical results showcase reductions of up to 22.6%, 27.8%, and 25.4% on the EMNIST, CIFAR-10, and Large Movie Review datasets, respectively, in comparison to FedAdp. Furthermore, when compared to FedAvg, FedImp achieves even more substantial reductions of up to 40.2%, 50.2%, and 16.4% on the same datasets. FedImp emerges as a promising solution, enhancing the efficiency of FL in scenarios with non-IID data distributions.

Student

*(Signature and full name)*

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1. INTRODUCTION

## 1.1 Problem Statement

In the era of data-driven decision-making, the role played by machine learning models has reached noticeability across a variety of applications. From delivering personalized recommendations to facilitating predictive analytics, these models have seamlessly integrated into the fabric of technological advancements. However, the traditional approach to model training, wherein data from diverse sources is aggregated in a singular location, has revealed itself to be fraught with considerable challenges. Privacy, security, and communication bandwidth arise as primary concerns within this centralized framework, reflecting the need for an innovative and adaptive solution. Amidst this backdrop, Federated Learning (FL) emerges as a transformative approach that holds the potential to overcome the aforementioned challenges and open a new era of decentralized machine learning model training [1]. This collaborative machine learning approach revolutionizes the traditional training methodology by enabling multiple devices to train a shared model without the necessity of exchanging raw data. Unlike the traditional centralized model training, wherein data is aggregated into a central server, FL harnesses the computational capabilities of local devices, ranging from smartphones to edge devices and other distributed nodes. This approach not only preserves the integrity of data privacy by ensuring sensitive information remains localized but also alleviates the need for extensive data transfers. The decentralized nature of FL renders it particularly suitable for applications in sectors where data confidentiality stands as a dominant concern, such as healthcare and finance. By mitigating the risks associated with centralized data aggregation, FL paves the way for a more secure and privacy-centric machine learning ecosystem. Furthermore, the flexibility in this approach aligns with the dynamic requirements of diverse sectors, offering a robust solution that overcomes the limitations imposed by traditional centralized training methodologies.

The FL process unfolds through a sequence of iterative model updates [2]. Initially, a local model update is computed on the basis of the distinct data repository that resides on the individual device. This local update encapsulates the local patterns and characteristics, reflecting the unique informational context of each device. Subsequently, these local model updates undergo an aggregation process, whereby they are amalgamated into a unified global model. The aggregation, often facilitated through algorithms like weighted averaging, ensures a cohesive integration of the diverse insights contributed by each device. This collaborative learning framework

is particularly noteworthy for its efficacy in allowing the model to gain insights from diverse data existing across the participating devices.

One of the crucial challenges faced by FL algorithms is the presence of non-Independently and Identically Distributed (non-IID) data across decentralized devices [3]. In traditional machine learning scenarios, IID assumptions are often made, implying that the training data is drawn from the same distribution across all participating devices. However, in FL, this assumption is frequently violated as devices may have diverse data distributions based on factors such as geographic location, user behavior, or device characteristics. Non-IID data poses a significant hurdle to the effective collaboration of decentralized devices in FL. When the training data on each device exhibits different statistical properties, models trained locally may specialize in capturing patterns unique to the local data distribution. Consequently, combining these locally trained models into a global model becomes challenging, leading to poor convergence rate and suboptimal generalization performance. Poor convergence arises due to the varying nature of locally trained models, leading to a difficulty in their aggregation into a unified global model. Suboptimal generalization performance further compounds the issue, as the resultant global model may struggle to capture insights from the diverse data distributions, rendering it less effective when applied to unseen data instances.

Effectively addressing the challenge of non-IID data in FL becomes a major goal in the proposed novel algorithms. Such algorithms must employ sophisticated strategies to diminish the impact of this challenge. The utilization of weighted aggregation stands out as a strategic approach, wherein the contributions of different devices to the global model are weighted based on factors such as data quality or relevance. Another strategy is the implementation of data augmentation techniques, which involves artificially enriching the diversity of local datasets. By introducing variations to the existing data, this approach allows the global model to develop a more comprehensive understanding of potential patterns and features, fostering adaptability and robustness. Additionally, the incorporation of adaptive learning rates becomes a reasonable approach, controlling the pace of model updates based on the specific characteristics of each device's data distribution. This adaptive mechanism enables a response to the varied complexities in non-IID scenarios, ensuring that the learning process is fine-tuned to the characteristics of each contributing device. This thesis introduces a novel algorithm designed to improve convergence rate of FL, particularly when confronted with the challenges posed by non-IID data distributions.