UQAC
Université du Québec
à Chicoutimi

BUILDING SAFE, SECURE AND TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR CRITICAL SYSTEMS

PAR ABDUL REHMAN

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES (M. SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

Artificial Intelligence (AI) has significantly transformed various aspects of our lives, yet concerns have arisen regarding security, trustworthy, reliability, and effectiveness, particularly in safety-critical systems. These systems can be biased, complex, and uncertain, leading to potential human or financial loss. An AI system should be robust, resilient, and transparent to ensure safety. The increasing adoption of these AI systems in critical domains such as healthcare, finance, and transportation has raised significant concerns about their decision's reliability, robustness, and trustworthiness. Various safety standards and principles are presented for critical systems. Current research focuses on improving AI systems, but little research has been done on ensuring safe AI. To develop and maintain safe AI, potential risks (including bias, data security, and susceptibility to external threats) must be identified, and procedures must be established to prevent and reduce them. Fairness, robustness, reliability, and AI alignment are the key criteria for Safe AI systems.

To ensure fairness and robustness, the development of AI models requires sufficient training data in terms of quantity and quality to make meaningful predictions. Current research focuses on training large models to make general decisions. However, training these models requires a lot of computational capacity and training data. To resolve the limited training data problem, researchers employ collaborative learning (aka Federated Learning), which collaborates with several smaller groups to pool their computational resources and local data and train a global model that benefits all participants. However, their reliability and safety are still a serious concern, especially in critical domains where errors or biases can have severe consequences.

The research begins with a conceptual framework of Safe AI, demystifying its principles and clarifying its role in fostering secure, robust, trustworthy, and ethically aligned AI systems. Within this context, collaborative learning emerges as a promising approach for safety-critical systems, but it faces two fundamental problems : (i) data heterogeneity across subgroups, which undermines fairness and robustness, and (ii) vulnerability to adversarial threats, particularly data poisoning, which compromises security and reliability.

To address these challenges, two novel contributions are presented. First, FedCIM introduces a mechanism to mitigate the effects of heterogeneous data distributions in collaborative learning, thereby enhancing fairness among participants and improving model robustness across diverse datasets. Second, RBFL proposes a reputation-based defence strategy that secures collaborative learning against data poisoning attacks, ensuring resilience and trustworthiness in training environments. Together, these contributions advance the state of the art in collaborative learning by embedding fairness, robustness, and security into its design. By integrating theoretical insights with practical solutions, this thesis demonstrates how the vision of Safe AI can be realized for safety-critical systems.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
SVM	Support Vector Machine
GAN	Generative Adversarial Network
FL	Federated Learning
DP	Differential Privacy
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
AI	Artificial Intelligence
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
NIST	National Institute of Standards and Technology
OECD	Organization for Economic Cooperation and Development
RL	Reinforcement Learning
SV	Shapley Value
EU	European Union
UK	United Kingdom
US	United States of America
OOD	Out of Distribution
DPD	Data Protection Directive
ISO	International Organization for Standards
IID	Independent and Identically Distributed
SGD	Stochastic Gradient Descent
API	Application Programming Interface
LLMs	large language models

DÉDICACE

To my mother and wife.

I dedicate this effort to you, with the aspiration that you will perpetually take pride in my endeavors.

I dedicate this work to all my family members for their encouragement, support, affection, and confidence in my academic pursuits.

To Professors Fehmi Jaafar and Darine Ameyed, who instilled confidence and faith in me and provided unwavering support during my studies.

To all individuals who, in various capacities, contributed to the finalization of this work.

May Allah bestow upon them health, prosperity, and His mercy.

REMERCIEMENTS

First, I want to express my gratitude towards GOD Almighty for granting me this chance, knowledge and strength to conduct this research study. May His peace and blessings be upon Prophet Mohammed (SAW), without the will and blessings of GOD, this achievement would not have been possible.

Without the assistance of many people, the research work discussed in this study would never have been possible. Firstly, I take this opportunity to express admiration and respect to my supervisor, Professor Fehmi Jaafar. He is an outstanding supervisor who provided me directions and yet has the freedom to explore various areas. Countless thanks to him because of my existing research support, guidance, compassion, motivation, and immense knowledge. I would like to sincerely thank my research co-supervisor, Professor Darine Ameyed, who had a significant impact on my research. Darine gives me constructive feedback on my several research works. All the meetings with them in preparing this thesis and research paper have been very supportive. Finally, my love and praise to my parents and wife for their diligence, support, and affection. Without them, all my accomplishments might never have been possible.

My thanks also go to all the staff of the Department of Mathematics and Computer Science (DIM) at UQAC.

Finally, thank you to my friends and loved ones for always helping me find a balance between my studies and personal life.

AVANT-PROPOS

This master's thesis represents the culmination of an intensive and rewarding academic journey, during which I focused on the study of safe, secure, and trustworthy artificial intelligence for critical systems. The published works listed below support my thesis proposition Rehman *et al.* (2026b,a, 2025a,b). Through advanced coursework and sustained research efforts, I developed strong foundations in artificial intelligence, cybersecurity, and distributed learning systems, with particular emphasis on the challenges posed by deploying AI in safety-critical and high-risk environments.

The primary objective of this work is to investigate how safe AI methodologies can be designed and integrated to mitigate key challenges such as fairness violations, lack of robustness, adversarial attacks, and security threats, particularly in collaborative and distributed learning settings (e.g., federated or multi-agent learning). This research explores techniques that enhance the reliability and trustworthiness of AI models while ensuring that their decisions remain fair, robust to perturbations, and resilient against malicious behaviour.

A central aspect of this thesis is the enhancement of fairness and analysis of robust and adversarially secure learning frameworks, along with mechanisms to detect, prevent, and mitigate attacks that target model integrity, data privacy, and collaborative training processes. By addressing these challenges, this work aims to contribute to the development of AI systems that can be safely deployed in critical domains, where errors, bias, or security breaches may lead to severe consequences.

This research journey has been marked by intellectual challenges, moments of uncertainty, and sustained effort, but also by meaningful progress and valuable insights. Each chapter of this thesis reflects a continuous process of exploration, experimentation, and critical thinking, driven by the goal of advancing trustworthy AI solutions for real-world, high-stakes applications.

I take full responsibility for the content and claims presented in this document. Acknowledging that scientific research is an evolving process, I welcome constructive feedback, comments, and suggestions to further strengthen and refine this work.

Each chapter of this thesis is based on my published research work :

- Rehman, Abdul & Ameyed, Darine & Jaafar, Fehmi. (2026). Safe and Intelligent Healthcare 5.0 : A Survey to Explore Challenges, Trends, and Innovative Solutions. ACM Computing Survey. (April 2026) (Conditional Accepted).

- Rehman, Abdul & Ameyed, Darine & Jaafar, Fehmi. (2026). Chapter 6 : Ensuring Safe and Robust AI : Fairness, Explainability, and System Reliability. Transparent

Intelligent : A guide to Explainable AI. Nova Science Publishers, Inc. (Accepted and Published).

- Rehman, Abdul & Ameyed, Darine & Jaafar, Fehmi. (2025). FedCIM : Handling Data Heterogeneity in Federated Learning to Improve Fairness and Robustness. 2025 IEEE Conference on Dependable, Autonomic, and Secure Computing (DASC) DOI : 98-107. 10.1109/DASC68382.2025.00020.

- A. Rehman, I. I. Mazu, F. Jaafar, D. Ameyed and H. B. Abdessalem, "RBFL : Securing Federated Learning against Data Poisoning Using a Reputation-based Approach," in 2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA), Doha, Qatar, 2025, pp. 1-10,
DOI : 10.1109/AICCSA66935.2025.11315401.

CHAPITRE I

INTRODUCTION

Artificial Intelligence (AI) is revolutionizing key sectors by enhancing productivity, facilitating personalized experiences, optimizing decision-making, and augmenting safety Salazar *et al.* (2026). AI is transforming numerous facets of our everyday existence and is being applied in almost every field. It automates repetitive tasks in sectors such as manufacturing and healthcare, enabling professionals to concentrate on more valuable work and enhancing total efficiency. AI improves customer engagement in e-commerce, marketing, and entertainment by providing personalized recommendations based on user behavior. Employing predictive analytics reveals trends within extensive datasets, assisting businesses and investors in making more informed, data-driven decisions Winther & Fredriksen (2025). The effective development and application of AI systems have increased exponentially over the past decade. However, with the proliferation of AI technology, many concerns have arisen regarding security, reliability, trustworthiness, and effectiveness, especially for safety-critical systems. Safety-critical systems are those in which a single mistake can have intolerable consequences, including a human's life or financial loss. The key challenge with the AI models is that these systems are biased, complex, overconfident, and uncertain about their decisions. Many crucial properties, such as lack of bias, resiliency, security, and transparency, are essential and may function as recommendations for successfully implementing safe AI systems. In addition, different standards and principles are presented for the function, safety, security, and privacy of AI systems. Essential standards for AI systems must be used in high-risk domains, such as medical devices and mechanized production units, which are mandated under the AI Act. To protect security, safety, and ethical concerns, these criteria address a variety of system life cycle characteristics. These

standards and principles must be fulfilled to evaluate AI systems and reduce risk. Therefore, the assurance of AI systems remains an open problem that needs to be addressed.

There are no universal definitions present for Safe AI. Various existing research studies Kearns & Ron (1999); Rehman *et al.* (2026a) explored safe AI in different ways. However, fairness, robustness, reliability, and accountability are the key characteristics that make the AI system safe. AI systems can be unfair due to various factors such as bias, data heterogeneity, limited training data, etc. Robustness is the property of an AI system by which a system must be resilient to uncertainties, errors, and security attacks. The complex nature of AI systems makes decisions ambiguous and unreliable. So, interoperability and transparency must be present in an AI system, especially for the critical domain. Much research is being done to improve various aspects of the AI system. However, not many studies have been presented to address the safety question. "What is SAFE AI, and how can we develop a safe cyber-physical and critical system ?" Therefore, resolving these issues is essential to delivering safe and secure AI-powered solutions for these safety-critical systems.

To mitigate the limited training data and the computational power, collaborative learning has been introduced that allows extensively distributed clients, such as mobile devices, hospitals, or financial institutions, to train a global model utilizing their local data while maintaining privacy. A centralized server aggregates the parameters of each client's local model to produce a global model and transmits this global model's parameters to each client for the subsequent training iteration Kairouz *et al.* (2021). The collaboration aims to develop a global model that outperforms individual local models trained by distributed clients. However, the data distribution across the clients differs significantly from the global distribution ; each client's local objective will be incompatible with the global optima, resulting in a drift in local updates. This data heterogeneity can lead to poor performance

across clients, as the global model tends to favor the clients representative of most data. Consequently, it induces biased performance for certain clients and raises critical concerns regarding the fairness and reliability of the global model.

Secondly, collaborative learning success relies on the continuous participation of diverse clients in training a global model. Malicious clients pose significant risks by attempting to poison the global model through manipulative activities such as providing incorrect or low-quality data. Additionally, market competition can hinder clients' participation in learning processes, such as developing credit score predictors for small businesses, as larger banks may be reluctant to share high-quality local data Yang *et al.* (2019). Furthermore, malicious clients or free riders can obtain the global model without participating in the training process, contributing random or adversarial updates that reduce model performance. Detecting these threats is crucial for maintaining the integrity and security of FL systems, especially in sensitive domains like healthcare Kumar *et al.* (2022), where accurate predictions are vital.

Several studies have been proposed to overcome the data heterogeneity problem motivated by the understanding that the global model may perform poorly among participant clients compared to a locally trained model. Existing approaches address this through various personalization techniques, including regularization, optimization Li *et al.* (2021a), multi-task learning Smith *et al.* (2017), transfer learning Luo *et al.* (2023), meta-learning Fallah *et al.* (2020), and knowledge distillation Zhu *et al.* (2021). The main aim of these diverse techniques is to generate personalized models for each client, disregarding the inherent uncertainties in model predictions, particularly for clients with significantly divergent data distributions. Additionally, these personalized methods are less suitable for clients with similar data distribution or less heterogeneous data. Recently, clustering-based techniques have been employed to mitigate data heterogeneity by training separate perso-

nalized models for clients having similar data distribution Ouyang *et al.* (2021). Existing approaches Xiao *et al.* (2021), Shu *et al.* (2023) rely on predefined cluster counts, which can only be determined with prior knowledge about data distributions. Since we do not know the actual distribution, extremely heterogeneous clients need more clusters, and slightly heterogeneous clients require fewer clusters. However, dynamic clustering has been proposed Sattler *et al.* (2019); Yan *et al.* (2024). Most of these approaches require all the clients to participate in each training round for efficient performance. Some studies cluster the clients based on the specific threshold Tian *et al.* (2022); Hsu *et al.* (2023). However, perfectly setting the threshold to obtain suitable clustering is challenging. Existing work only focuses on clustering clients with different approaches ; very little literature focuses on the fairness and robustness of the approach.

Similarly, various solutions have been presented to address adversarial robustness and poisoning in the FL process, including principal component analysis & defense mechanism Khraisat *et al.* (2025); Zhao *et al.* (2024), Shapley additive explanation & support vector machines Khuu *et al.* (2024), local model test voting, pullback-Leibler divergence anomaly detection ‘Li *et al.* (2022), and data augmentation techniques Yin & Zeng (2024). However, these solutions require multiple retrainings, are computationally costly, and are challenging to apply in practice.

1.1 RESEARCH QUESTIONS

The following research questions serve as the basis for this work :

- What is SAFE AI in general terms ? What are the fundamental principles that render AI systems SAFE, designed to ensure safety, reliability, and ethical compliance while addressing emerging challenges in safety-critical systems ?

- How can collaborative learning frameworks be improved to address data heterogeneity while enhancing fairness and robustness for safety-critical applications ?
- How can reputation-based techniques enhance the resilience of collaborative learning against adversarial risks like data poisoning, hence securing safety-critical AI systems ?

1.2 THESIS CONTRIBUTIONS

This thesis integrates three key research contributions : a conceptual exploration of safe AI Rehman *et al.* (2026a), along with my two original research studies Rehman *et al.* (2025a,a) that address fairness, robustness, and security in collaborative learning for safety-critical systems. The main narrative is unified by the core research problem : how to design, develop, and deploy AI-driven systems that are not only intelligent and efficient but also safe, fair, robust, and trustworthy. This thesis integrates these contributions into a unified story, delineating the progression from fundamental notions to methodological advancements and practical ramifications for the future of safe AI. The three main contributions of this thesis are as follows :

Conceptual Framework of SAFE AI

We provide the conceptual framework of SAFE AI Rehman *et al.* (2026a) that emphasizes the significance of safe AI, particularly for safety-critical systems. We define safe AI and its actual significance. We show various available safety standards. To ensure safety and reliability, we outline the different factors that should be incorporated into any AI system. Fairness, robustness, reliability, and AI alignment are the essential characteristics of Safe AI systems. We discussed the significance, taxonomy, and problems associated with these factors. We also examine the various measures implemented by governments to safeguard the safety of AI.

Handling Data Heterogeneity in Collaborative Learning

To mitigate the current research problem of data heterogeneity and provide a more efficient solution to tackle the fairness and robustness of collaborative learning, we propose Rehman *et al.* (2025a) an efficient approach to handling non-IID data in FL, eliminating constraints such as predefined cluster counts, client participation, and inferential datasets. We propose an improved dynamic clustering approach that iteratively groups clients with similar data distributions, treats each cluster as a separate task, and applies multi-task learning to provide personalized models for each task. By recognizing each client's data distribution as a unique task, our method aims to preserve the benefits of collaborative learning while addressing the diversity in client data.

Reputation Awareness Contribution Evaluation framework

We propose Rehman *et al.* (2025a) a reputation-aware, fair contribution evaluation, and adversarial robustness approach that utilizes the cosine gradient as a utility function. To mitigate the costly computation of Shapley Value (SV), we employed a fast approximation of SV called TMCShapley Liu *et al.* (2021) to estimate the marginal contribution of each client. Instead of retraining every permutation of clients, it utilizes prior gradient updates from all clients to reconstruct the FL sub-model, which improves the computation cost while closely approximating the actual SV value. Additionally, inspired by the research Kang *et al.* (2019); Xu & Lyu (2020), a history of gradient updates from every server client managed by the server is tracked to ensure that clients consistently contribute positively to the global model's performance over multiple training rounds. Considering the sensitivity of fair contribution evaluation and adversarial robustness for critical domains, we also use the healthcare domain dataset, as most of the existing research work requires a test set to measure the performance of a global model, raising privacy concerns.

1.3 ORGANIZATION OF THESIS

The remainder of the thesis is organized as follows :

- Chapter 2 Literature Review—discusses the pertinent literature and the background knowledge needed to understand our proposed approaches discussed in this thesis efficiently. (This chapter is based on my publication Rehman *et al.* (2026b))
- Chapter 3 presents a conceptual framework that signifies the importance of a safe AI system, delineates the concept of safe AI, and demystifies the different attributes of safe AI. (This chapter is based on my publication Rehman *et al.* (2026a))
- Chapter 4 shows an approach, FEDCIM, that handles data heterogeneity in federated learning to improve fairness and robustness for safety-critical domains. (This chapter is based on my publication Rehman *et al.* (2025a))
- Chapter 5 presents a reputation-aware contribution evaluation approach that secures federated learning against adversarial robustness and free riders. (This chapter is based on my publication Rehman *et al.* (2025b))
- Chapter 6 Conclude the thesis and provide future directions.

CHAPITRE II

PRELIMINARIES AND RELATED WORK

2.1 PRELIMINARIES

This section outlines the essential preliminaries necessary for a comprehensive understanding of our methodology. It commences with an overview of collaborative learning, then presents several concepts such as fairness, robustness, cosine similarity, and Shapley value. Next, it provided relevant research work that had previously been proposed in a similar domain to ours.

2.1.1 FEDERATED LEARNING

Federated Learning (FL) Lyu *et al.* (2020) is a type of collaborative learning that allows extensively distributed clients to train a global model utilizing their local data while maintaining privacy. A centralized server aggregates the parameters of each client's local model to produce a global model and transmits this global model's parameters to each client for the subsequent training iteration ; a standard FL system comprises a set of honest clients $N = \{1,..,n\}$ and a trusted central server. The objective is to learn a global model collaboratively without exchanging the local dataset by minimizing a global loss function $F(w)$. The global loss function is the weighted average of the local loss functions of N participants.

$$minF(w) = \sum_{k=1}^{N} p_k F_k(w) \tag{2.1}$$

Where $F_k(.)$ is the local loss function for a client $k\ \varepsilon\ N$ with a local dataset D_k and $n_k = |D_k|$. Here p_k is the weight of the client k such that $p_k >= 0$ and $\sum_{k=1}^{n} p_k = 1$.

$$F_k(w) = \sum_{j_k=1}^{n_k} f_{j_k}(w; x^{j_k}, y^{j_k}) \tag{2.2}$$

We consider the widely used optimized version of the FL algorithm, Federated Averaging (FedAvg McMahan *et al.* (2016)), which is based on the simple aggregation strategy of iterative model averaging.

2.1.2 COSINE SIMILARITY

Cosine similarity is a mathematical metric that calculates pairwise similarity between two data points Lahitani *et al.* (2016). In ML, it is widely utilized to measure the alignment between the gradients of two models by estimating the similarity between two non-zero vectors in an inner product space. It measures the cosine angle between the vectors to ascertain the direction, whether pointing in the same or opposite direction. For two gradients ∇w_c and ∇w_N the cosine similarity is

$$Cosine(\nabla w_c, \nabla w_N) = \frac{\langle \nabla w_c, \nabla w_N \rangle}{\|\nabla w_c\| \|\nabla w_N\|} \tag{2.3}$$

It returns a scalar value range $[-1, 1]$ where a value closer to 1 signifies increased similarity and a value closer to -1 indicates greater opposition between the two gradients.

2.1.3 SHAPLEY VALUE

The Shapley value (SV) Shapley (1953) is a fairness concept in data valuation, originating from cooperative game theory. It distributes the collaborative task payoff among agents. The fairness criteria of the SV play an integral role in ensuring a fair payoff and valuation by satisfying the following properties.

— **Symmetry** : For any two clients $i, j \in [N]$, if for any subset of clients $S \subseteq [N] \setminus \{i, j\}, U(S \cup \{i\}) = U(S \cup \{j\})$, then $v(i) = v(j)$.

— **Zero element** : For any client $i \in [N]$, if for any subset of clients $S \subseteq [N] \setminus \{i\}, U(S \cup \{i\}) = U(S)$, then $v(i) = 0$.

— **Additivity** : If the utility function U can be expressed as the sum of separate utility functions, namely $U = U_1 + U_2$ for some $U_1, U_2 : 2^I \to \mathbb{R}$, then for any client $i \in [N]$, $v(i) = v_1(i) + v_2(i)$, where v_1 and v_2 are the evaluation metrics associated with the utility functions U_1 and U_2, respectively.

— **Efficiency** : $v([N]) = \sum_{i \in [N]} v(i)$

The metathetical representation of the Shapley value can be written as follows :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

Here $\sum_{S \subseteq N \setminus \{i\}}$ represents the sum of all possible coalitions that a player i can join, which $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ denotes the weighted average (probability that the player makes some contribution) and $(v(S \cup \{i\}) - v(S))$ provides the marginal contribution of the client i to the coalition S.

The symmetry principle ensures fair assessment of each client's contribution, while the zero-element property prevents clients from being considered if they aren't beneficial to

the coalition. Additivity and efficiency properties ensure the utility of the coalition equals all clients' contributions. Although SV ensures fairness, it requires repeated training with different permutations of FL participants, which is computationally expensive and not applicable to multiple client scenarios.

In the next two subsections, we define the concepts of robustness and fairness in machine learning based on previous work Fallah *et al.* (2020); Zhu *et al.* (2024a). Our work focuses on the unfairness issue caused by the non-IID data distribution among clients. This type of fairness is called representation disparity, a serious challenge in federated networks that can cause significant performance variations across clients, potentially leading to uneven outcomes and, thus, a decrease in the robustness of FL models.

DEFINITION 1 (FAIRNESS) :

Fairness is the uniformity of model performance distribution among users. For any two models, w_1 and w_2, we can say that model w_1 is more fair than w_2 if the test performance distribution of the model w_1 across the network exhibits greater uniformity than that of w_2. It can be denoted as $\text{var}\{F_k(w_1)\}_{k\in[K]} < \text{var}\{F_k(w_2)\}_{k\in[K]}$, where $F_k(.)$ represents the test loss on client $k \in [K]$, and *var* signifies the variance of the accuracy Fallah *et al.* (2020).

2.1.4 DEFINITION 2 (ROBUSTNESS) :

We delineate the ideas of robustness according to prior research Li *et al.* (2021b); Zhu *et al.* (2024b). The current literature has examined various forms of training attacks Lyu *et al.* (2022). However, we concentrate on poisoning, Byzantine attacks, and free-riders, wherein malicious clients might transmit random or malicious updates to the server to compromise training. The precise notion of robustness is : For any two models, w_1 and

w_2, we can say that model w_1 is considered more robust than model w_2 if the average test accuracy for benign clients utilizing model w_1 exceeds that of benign clients employing model w_2 under adversarial conditions.

2.2 RELATED WORK

Safe AI currently lacks a universal definition. Various existing research studies Kearns & Ron (1999); Rehman *et al.* (2026a) explored safe AI in different ways. However, fairness, robustness, reliability, AI alignment, and accountability are the key characteristics that make the AI system safe. Building on this foundation, this thesis has examined fairness and robustness as two core pillars for constructing safe AI systems, demonstrating how bias mitigation techniques and adversarial resilience mechanisms can be systematically embedded in AI pipelines for safety-critical applications. Fairness-focused research has largely concentrated on detecting and correcting disparate treatment across protected groups, while robustness research has addressed a model's resilience to perturbations, distributional shift, and adversarial manipulation. Together, these two pillars represent a substantial step toward operationalizing the abstract principles proposed by frameworks such as NIST's AI principles & features of Standards & (NIST,N) and the OECD AI Principles for Economic Cooperation & (OCED), translating high-level trustworthiness attributes into measurable, testable properties.

However, existing literature tends to treat these pillars in isolation, evaluating fairness or robustness as standalone objectives rather than as interdependent components of a unified safety framework. This siloed treatment obscures important trade-offs : interventions that improve robustness (e.g., adversarial training) can inadvertently degrade fairness across subgroups, and fairness-constrained optimization can, in turn, reduce a model's resilience to distributional shift. Moreover, the broader body of work reviewed in this thesis reveals that

transparency, accountability, and explainability are frequently addressed as afterthoughts or compliance checkboxes rather than as design-time considerations, leaving a gap between regulatory expectations (e.g., GDPR, HIPAA) refer (b); Government (2024); (GDPR); Goddard (2017) and the technical mechanisms needed to satisfy them in practice.

Perhaps most critically, AI alignment — ensuring that a system's learned objectives and behaviours remain consistent with human intent, especially as models scale in capability and autonomy — remains conspicuously underexplored relative to fairness and robustness in the safety-critical systems literature. Existing standards and taxonomies (Section 3.3) Fig 3.1) acknowledge alignment only implicitly, if at all, despite its centrality to preventing unintended or emergent behaviors in complex AI systems. This omission motivates the approach proposed in this thesis : rather than advancing fairness and robustness as isolated pillars, we position them alongside transparency, accountability, and most critically — AI alignment, within a single integrated framework, allowing the interactions and trade-offs between these properties to be assessed jointly rather than in disconnected silos.

2.2.1 FEDCIM Rehman *et al.* (2025a)

Several approaches have been presented to tackle the challenges of data heterogeneity in FL. For example, Sattler et al. Sattler *et al.* (2019, 2021) proposed the first cluster FL approach, which recursively bi-partitions the two client groups based on the diversity of the gradient update. Ghosh et al. Ghosh *et al.* (2020) presented an L2 distance-based clustering technique that utilizes a weight-sharing mechanism to train separate final layers for each cluster. This approach requires continuous communication and is significantly influenced by the initial configuration of the cluster count. The study Gong *et al.* (2024) proposes a clustering method that uses local training epoch adjustment and weighted voting to filter non-IID data efficiently, which may raise privacy concerns.

Different strategies have been investigated for dynamically clustering clients. Liu et al. Liu *et al.* (2024) presented a secure aggregation method that partitions users into clusters using gradient sparsification and encryption mechanisms to reduce communication overhead and eliminate semi-honest customers. Some existing studies Yan *et al.* (2024); Morafah *et al.* (2022) utilize an adjacency matrix and incremental clustering to cluster clients based on data distribution similarity but require an inference dataset to measure similarity. Similarly, threshold-based approaches Tian *et al.* (2022); Hsu *et al.* (2023); Li *et al.* (2023); Du *et al.* (2023) aim to eliminate a fixed number of clustering requirements using a similarity matrix, but accurately adjusting thresholds to avoid losing valuable clients remains a challenge.

Duan et al. Duan *et al.* (2022) propose flexible clustering based on k-means similarities in model parameters, balancing communication and accuracy. Holzer et al.Holzer *et al.* (2023) use dynamic k-means clustering, addressing privacy and security concerns by dynamically prioritizing updates during clustering. However, sharing each cluster's information with the server may raise privacy and security challenges. Research in Lai *et al.* (2024); Tan *et al.* (2022) addresses communication challenges in clustering clients by analyzing prototype similarity. Every client must compute and share the data prototype with the server, which raises privacy concerns. Similarly, Lu *et al.* (2024) clusters clients based on client data similarity and predictive uncertainty. Predictive uncertainty raises privacy concerns and increases computational costs due to data similarity estimation, necessitating each client to download the global model from the server.

GAN-based clustering Kim *et al.* (2020) compresses data into latent space, improving accuracy and privacy. However, diverse latent representations may result in multiple clusters. A similar study Radovič & Pejović (2023) clusters clients without requiring training or data labels, using autoencoders to create a latent space and find the similarity based on

correlations between client embeddings, but it still requires unlabeled data collection. The study Ouyang *et al.* (2021) proposes an optimization method to improve communication costs during client clustering, utilizing correlation-based client selection and dropping out of straggler clients but requiring client participation in all communication rounds. The studies in Xiao *et al.* (2021); Shu *et al.* (2023) propose a method to measure client similarity using local model parameters with a predefined number of clusters initially; still, this method faces challenges, as insufficient or excessive clusters can lead to incorrect clustering of clients.

Research on clustering clients primarily focuses on the efficiency of clustering and communication overhead, with less emphasis on fairness and robustness. Hierarchical-based clusteringLi *et al.* (2023), proposed in a human activity recognition setting, faces challenges in setting the right threshold for optimal clustering. The larger the clustering threshold, the more difficult it is to merge the two clusters. To tackle the challenges, including the predefined number of clusters and the selection of all clients in every training round, robustness, and fairness, we presented an iterative strategy that iteratively clustered clients by reassigning them to a new cluster. This leads to outlier detection and enhances fairness and robustness.

2.2.2 RBFL Rehman *et al.* (2025b)

Recent research has tackled the poisoning challenges in the FL process. Some existing studies identified malicious participants by analyzing parameter updates in vulnerable training rounds using PCA Khraisat *et al.* (2025) and the defense mechanism Zhao *et al.* (2024) by grouping clients based on their model updates. Khuu et al.Khuu *et al.* (2024) employed SHAP values and SVM to differentiate malicious from honest clients. To defend against label-flipping attacks, Li et al. Li *et al.* (2022) utilized the local model test voting and

pullback-Leibler divergence to detect anomaly parameters. Similarly, the data augmentation technique is presented to defend against data poisoning attacks in data heterogeneity situationsYin & Zeng (2024).

Various research studies have been conducted to evaluate clients' contributions in FLWu *et al.* (2024). Here are two notions we must consider, i.e., Utility functions v calculate the importance and usefulness of coalitions, which will be used by valuation functions ϕ to output the final contribution depending on the existence of other clients' data in FL. Different utility and valuation function designs yield data valuations for various properties. For instance, research in Xue *et al.* (2021); Wang *et al.* (2019) uses the influence function to estimate client influence on global models, excluding specific clients and re-training to compare results to calculate the impact, potentially raising fairness concerns. Similarly, Xue et al. Xue *et al.* (2021) propose an alternative methodology to measure client influence, excluding updates during aggregation in training rounds. However, this approach may create unfairness, as it doesn't affect the global model's performance if two clients have similar influences.

Other existing approaches Ghorbani & Zou (2020); Li *et al.* (2024); Ghorbani & Zou (2019); Jia *et al.* (2019); Chen *et al.* (2020); Lyu *et al.* (2020) depend on the global model's validation accuracy to calculate the performance. In most FL scenarios, the aggregator server has no test data to measure performance. Sun et al. Sun *et al.* (2023) proposed a correlation and data volume-based contribution estimation approach. They measure the participant's data quality by combining the data quantity and correlation scores. However, a large dataset is not the data quality measuring criterion because large data can have redundant data, and a small data volume can have high-quality data. In contrast with influence-based methodology, Sim et al. Sim *et al.* (2020) proposed a fair incentive mechanism based on client contribution, quantifying model quality by reducing uncertainty before and after

training on local data. For the marginal contribution estimation, most of the studies Wang *et al.* (2020); Shapley (1953); Chen *et al.* (2024) employed SV, providing fairness ; however, it is computationally costly and challenging to apply in practice due to multiple client permutations.

Collaboration-based approaches use data utility functions to quantify marginal contributions. A robust volume-based matrix Xu *et al.* (2021b) is proposed as an alternative to validation accuracy-based approaches. It takes robust volume to measure the data utility and Shapley value for estimating the marginal contribution. In the same context, Xu et al. Xu *et al.* (2021a) presented their study based on the gradient similarity as a utility function and Shapley value for the marginal contribution. Several studies Shi *et al.* (2021); Zhang *et al.* (2024); Ghosh *et al.* (2024); Siomos & Passerat-Palmbach (2023) utilize the different variants of Shapley value as data valuation matrices because of their fairness properties. However, due to its computationally expensive nature, it is impractical to apply it in real cases when the number of clients is large. To mitigate the shapley expensive retraining computation, Liu et al. Liu *et al.* (2022) employed the GTGShapley approximation. It involves creating numerous aggregated FL sub-models, which incorporate local model updates from various combinations of FL participants. However, it did not provide robustness and security for their proposition.

2.3 CONCLUSION

This chapter describes the fundamental concepts required for a thorough comprehension of our methodology. The chapter begins with an overview of collaborative learning, followed by the introduction of concepts including Fairness, Robustness, Cosine Similarity, and Shapley Value. Subsequently, it presented pertinent research studies that had been

previously proposed in a comparable field to ours. The subsequent sections elaborate on these notions to explain the experimental technique and the set of methods studied.

CHAPITRE III

CONCEPTUAL FRAMEWORK OF SAFE AI

This chapter (is based on my publication Rehman *et al.* (2026a)) will elucidate the significance of a safe AI system and its importance and desirability for safety-critical applications. Subsequently, we stated the Conceptual framework of safe AI, safety standards, the guiding principles of prominent organizations, and the various fundamental pillars that should be incorporated into a Safe AI system.

3.1 WHY IS SAFE AI IMPORTANT ?

AI has revolutionized our daily lives, encompassing smart homes, healthcare, finance, industry, and beyond Rehman *et al.* (2026b). In simple terms, AI can be stated as an automatic system that can fabricate instinctive predictions for a given set of defined objectives by humans. Although AI systems have proven to be incredibly effective in various industries. However, there are numerous concerns about these AI systems' safety and security, especially for critical systems. Any tiny fault or malfunction can risk substantial loss of life, assets, environmental damage, or privacy. Some examples of critical systems include healthcare, autonomous vehicles, and automatic plants etc. The ultimate objective of safe AI is to control or at least overcome the polarity between safe and complex AI systems' improbable behavior. "Safe AI" can be defined as a system that can be operated with a very low risk of failure and can minimize the damage experienced from unsolicited repercussions. In a safe environment, solutions to deal with the implications of malfunction are typically devised where areas of possible failure are identified. Let's take the example of autonomous vehicles ; in an exceptional case where one wheel does come off, the rest of the system should, therefore, make sure that the vehicle slows down fairly straight and

comes to a safe halt. For an airplane case, if one engine fails, it should also be able to continue flying on the remaining engines in the unfortunate scenario. The aircraft should be able to smoothly glide to a safe landing if it only has one engine (for a fighter jet case, the pilot is provided an ejector seat). While it's frequently debatable if this method of designing systems is always effective, every effort is typically taken to create engineered systems that are as safe as possible. Therefore, in these scenarios, an AI system should have a safe mechanism to at least overcome the exceptional error or malfunction that occurs in a system. This can be even worse for the safety-critical systems and cyber-physical systems. These limitations of an AI system deter the widespread adoption of AI systems for safety-critical domains. Hence, the implementation of Safe AI is indispensable to increase the user's trust and confidence in the system. In the next section, we discuss the different critical infrastructures & standards and how safe AI is necessary for these critical systems to mitigate the associated risk.

3.2 SAFE AI AND CRITICAL SYSTEMS

U.S. Cybersecurity Infrastructure and Security Agency defines 16 critical infrastructure sectors : chemical, commercial facilities, communications, critical manufacturing, dams, the defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, health care, information technology, nuclear reactors, transportation systems, and water and wastewater systems refer (a). Other countries also define critical infrastructure sectors as generally the same and universal.

In recent years, the application of AI in critical systems has surged, enhancing patient outcomes in healthcare, optimizing transportation through self-driving cars, and improving energy efficiency in electricity generation. AI assists medical professionals by analyzing patient data, while in manufacturing, it helps monitor production processes for efficiency.

Besides the many advantages of AI in critical systems, it comes with several potential risks and challenges. As AI technologies advance and become increasingly incorporated into diverse facets of our daily routines, ensuring AI safety is essential. The advancement of AI capabilities generates concerns about unintended consequences and potential exploitation. If not sufficiently developed, monitored, or regulated, it may result in significant errors, propagate biases, or be exploited for malicious intent. It is essential to identify the possible glitches in the system, continually monitor and conduct risk assessments, and manage these accordingly. It's crucial to emphasize the clear distinction between safety and reliability. Reliability concerns the frequency of system failures, while safety focuses on the consequences of such failures. However, strong correlations exist between these concepts. To ensure reliability and safety, developers must comprehend the outcomes, encompassing the systems and their components under all circumstances. Adhering to various best practices and ethical considerations is imperative to mitigate potential risks and challenges associated with deploying AI in critical systems.

Many ethical and human concerns related to AI arise from weak security measures, model malfunctions, unauthorized data access, and integrity issues. Various countries, including Australia, Canada, Turkey, the UK, the US, New Zealand, and the Netherlands, have passed privacy legislation Government (2024). For example, the US government passed the Health Insurance Portability and Accountability Act (HIPAA) in 1996 to enhance the security of healthcare data refer (b). Similarly, the General Data Protection Regulation (GDPR) was implemented in 2018 to replace the Data Protection Directive (DPD) in the European Union (EU) (GDPR). The GDPR introduced extensive laws for data protection and privacy Goddard (2017). These regulations empower individuals ; however, the effectiveness of these regulations varies. There is a pressing need for standardized AI safety protocols, as current international rules provide limited guidance for assuring AI-based

safety, despite proposed principles from various organizations aimed at managing AI risks. There are specific international rules and standards defined for safe critical systems (AI & non-AI), which can be seen in the Table. 3.1.

TABLEAU 3.1 : Safety Standards across various domains, sectors, and system types

Domain	**Sector**	**Safety Standard**			
		Functional Safety	**Heteronomous**	**Autonomous**	**AI Standards for Safety Systems**
Transportation	Space	ECSS-Q-ST-30C/40C			
	Railway	EN 5012x	IEC 62290, IEC 62267		
	Avionics	ARP 4754, DO-178C, DO-254	ASTM F3269-21		(ARP 6983)
	Automotive	ISO 26262	ISO/PAS 21448	ISO 4804, ISO 5083, UL 4600	(ISO/AWI PAS 8800)
Industrial	Robotics	ISO 10218-1			
	Mining & Earth Moving Machinery	ISO 19014	ISO 17757, ISO 16001, ISO 18758-2		
	Industrial Machinery	ISO 13849-1	ISO/TR 22100-5, ISO 3691-4		
	Agriculture	ISO 25119	ISO 10975, ISO 14897		
Generic		IEC 61508	VDE-AR-E 2842-61		(ISO 5469)

Next, we discussed the taxonomy of SAFE AI in light of different organizations' principles and attributes. Then, we elaborate on the important pillars in detail.

3.3 TAXONOMY OF SAFE AI

3.3.1 WHAT IS SAFE AI IN ACTUAL ?

The term "SAFE AI" lacks a standardized definition in AI but encompasses various interpretations linked to AI risk management frameworks. The National Institute of Standards and Technology (NIST) has established eight features and principles for trustworthy AI systems of Standards & (NIST,N), while the OECD lists four essential principles for risk mitigationfor Economic Cooperation & (OCED). US Executive Order 13960 requires reliable AI systems based on specific principles. The UK's recent report on AI safety identifies risks and suggests mitigation strategies, and the EU's Ethics Guidelines for Trustworthy AI provide seven fundamental principles.

Category of Risk	NIST Taxonomy	OECD AI Principal	EU AI Attributes	US EO 13960	UK Risk Mitigation Approaches
Technical Design Attributes	Accuracy Reliability Robustness Security & Resilience	Robustness Security	Technical-Robustness	Purposeful and Performance-driven Accurate, Reliable, and Effective Secure and Resilient	Training of Trustworthy model Mitigating Biases & Improving Fairness
Guiding Principals for Trust-worthiness	Fairness Accountability Transparency	Traceability to Human value Transparency and Responsible Accountability	Human Agency and oversight Data governance Transparency Diversity and fairness Environmental and Societal well-being Accountability	Lawful and Respectful of our nation's values Responsible and Traceable Regularly monitored Transparent Accountable	Privacy Security
Socio-technical Attributes	Privacy Safety Explainability Interpretability Absence of Bias	Safety	Safety Privacy Non-discrimination	Safe Understandable by subject matter, experts, users, and others	Risk Assessment & Management Monitoring & Intervention

FIGURE 3.1 : Taxonomy of Safe AI proposed by different Organizations

In light of these taxonomies, principles, and properties (which can be seen in Fig 3.1), we define the interpretation of SAFE AI, which states that an AI solution must be robust, fair, secure, reliable, accountable, address ethical concerns, and be explainable to human users. These factors must be considered to build a safe AI system, especially for safety-critical domains.

The rest of the chapter will discuss the four most important pillars to make the system safe, i.e., fairness, robustness, transparency & accountability, and the important one that AI researchers neglect is AI alignment.

3.3.2 FAIRNESS

Fairness in AI seeks to reduce biases and discrimination, ensuring equal treatment for individuals or groups across diverse social, cultural, and demographic parameters. It is a fundamental principle of AI development aimed at enhancing social justice and equality while mitigating harm to vulnerable or marginalized groups Caton & Haas (2024). Fairness can be divided into two categories (see Fig. 3.2).

According to **individual fairness**, AI systems must treat similar individuals equally, irrespective of their specific traits. It emphasizes tailored treatment based on individual characteristics and merits rather than group membership. In contrast, **Group Fairness** aims to guarantee that AI systems handle various demographic groupings (such as race, gender, and age) equally. It seeks to provide equitable opportunities and results for all groups while preventing discrimination or differential effects against certain groups. For example, let's take an example of bank data in which the demographics of its customers will vary depending on the bank branch (formed by sensitive attribute(s), such as gender). An unfair system in which the model performs well for the male subpopulation but cannot make an

appropriate decision for the female subpopulation. Group fairness can be further divided into the following categories. **Statistical (demographical) Parity** entails that a model's prediction is independent of the subgroup (gender, race, etc.). In other words, the distribution of outcomes predicted by an AI system must be constant across all demographics. It seeks to guarantee that decisions are made without considering protected characteristics (such as gender or ethnicity) that can produce biased results. **Equalized Odd** ensure that the prediction performance of an AI system (such as accuracy, false positive rate, and false negative rate) must be similar across all demographic groups. It aims to preserve overall accuracy while reducing the differences in prediction errors between groups.

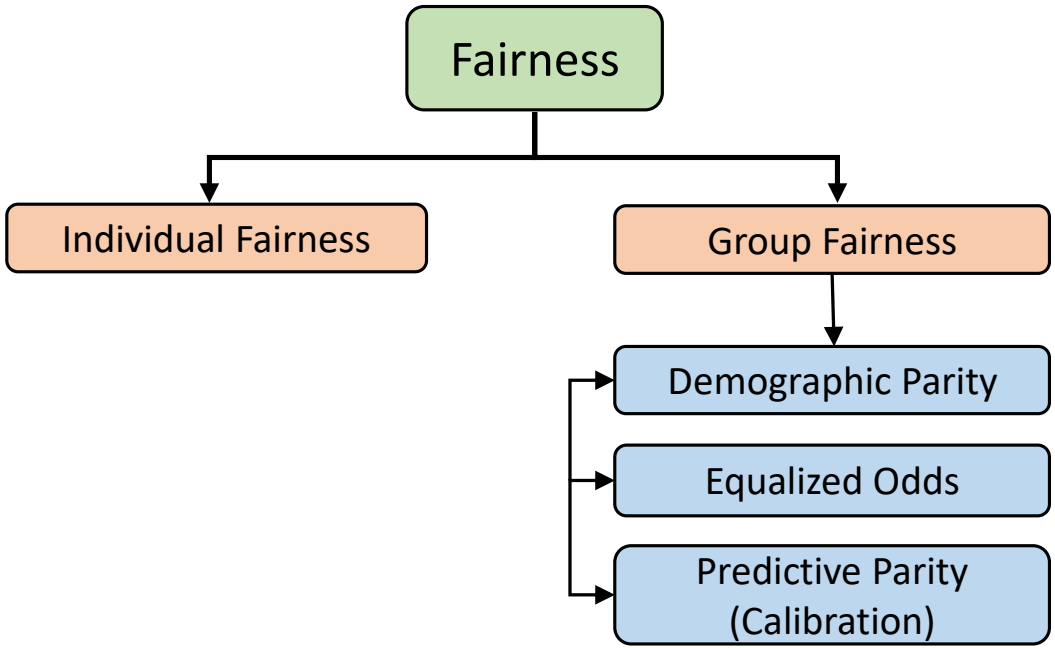

FIGURE 3.2 : Types of Fairness

To achieve **Predictive Parity (calibration)**, an AI system's probability of predicting positive results (true label) must be constant across demographic groupings, independent of other factors. It attempts to stop decision-making biases that could favor some groups over others.

Research on algorithmic and procedural fairness has introduced dynamic definitions of fairness. Algorithmic fairness concerns the standards for AI algorithms, focusing on identifying and mitigating biases to ensure equitable outcomes. Procedural fairness addresses the fairness of the methods employed in AI development and decision-making, highlighting the importance of transparency, accountability, and inclusivity in the design and governance of AI systems to maintain fairness principles.

3.3.3 ROBUSTNESS

We define robustness in two key terms, namely security and robustness. Security involves potential attacks that could induce an AI system to perform as intended or to compromise its functionality, either through system takeover or data exploitation. In contrast, robustness is defined as the AI system's capability to operate reliably even under these attacks or unintended behaviors. It necessitates the implementation of protective measures to safeguard AI models, datasets, and infrastructure from unauthorized access and misuse Qayyum A (2021). Robustness can be further categorized based on various attack types and the unintentional actions of the AI system (also shown in Fig. 3.3).

The goal of **Adversarial Robustness** is to protect AI systems against adversarial attacks, where malicious actors deliberately manipulate input data to mislead the system or induce errors in its predictions. Robust AI models must resist these attacks, functioning accurately and consistently even when confronted with hostile inputs. **Input Perturbation Robustness** ensures AI systems stay consistent and stable in the presence of noisy, distorted, or corrupted input data. Robust models should be able to adapt well to changes in the input data distribution, reducing the adverse effects of noise or uncertainty on their performance. **Model Robustness** describes the generalization ability and stability of AI models across diverse tasks, domains, or utilization scenarios. Robust models avoid overfitting to particular

datasets and perform consistently & reliably under various conditions and scenarios. **System Robustness** incorporates the overall resilience of an AI system, which includes data pipelines, inference processes, model training techniques, and deployment infrastructure. Robust AI systems must be engineered to tolerate faults, malfunctions, and disturbances, guaranteeing consistent and dependable functioning in practical applications.

Another category is the robustness against free riders and Byzantine clients. **Byzantine attackers, or free riders,** are those clients in centralized ML that do not contribute to training a model ; instead, they influence training and get an updated final global model without contributing. They can disrupt the convergence and overall performance of the global model by sending random gradients to a central server. Additionally, because the server cannot access the clients' training data or keep track of their local training procedures, it is difficult for the server to identify the Byzantine clients.

3.3.4 TRANSPARENCY & EXPLAINABLILITY

In the context of Safe AI, transparency refers to the clarity, accountability, explainability, and trustworthiness of AI systems or algorithms. It provides users or stakeholders insight into how AI decisions are made and the factors influencing outcomes. It enhances trustworthiness and accountability and promotes the responsible use of AI in critical systems Siala & Wang (2022).

Transparency involves revealing AI algorithms' logic, functionality, and inner workings so that users and stakeholders can comprehend how these systems analyze data, generate predictions, and make decisions. Transparent algorithms make it easier to examine, validate, and hold people accountable, which increases people's confidence in AI systems. To promote accountability and fairness in AI research, transparent data practices entail

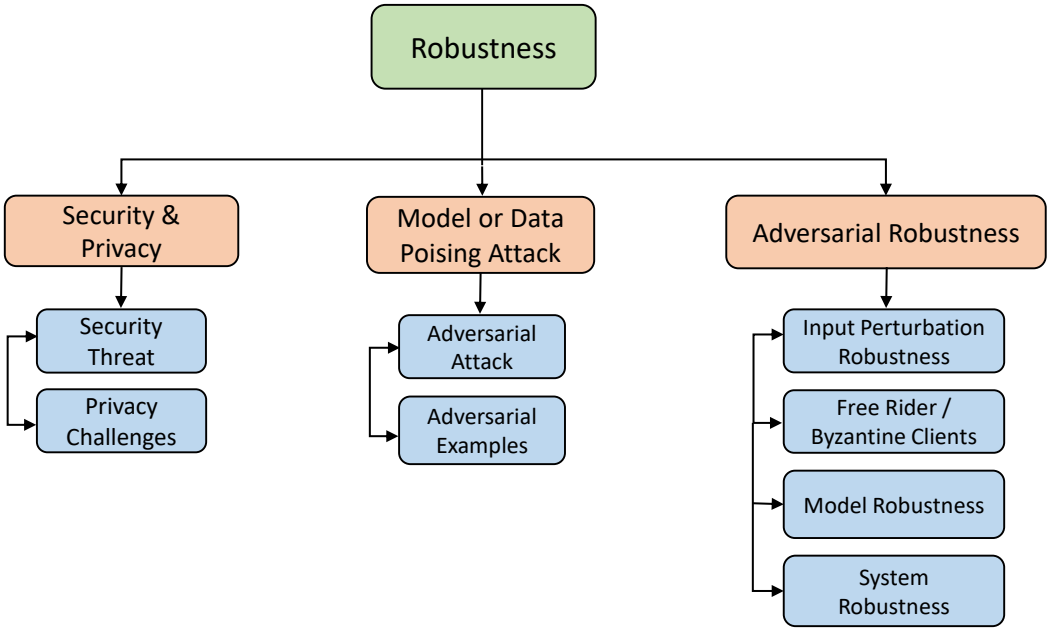

FIGURE 3.3 : Various Forms of Robustness

identifying data sources, collection methodology, processing processes, and potential biases or restrictions. Accountability entails developing procedures for tracking and explaining AI decisions, guaranteeing accountability for results, and providing recourse for persons impacted by AI systems. The concepts of explainability and interpretability are closely related. Explainability encompasses a broader context than interpretability. It refers to translating or explaining an ML model's complex behavior (black boxes) into human-understandable terms. On the other hand, interpretability examines the workings of the inner components of the AI/ML models to fully comprehend the reasons and methods behind the model's predictions. By bridging the knowledge gap between sophisticated AI models and human comprehension, XAI techniques enable users to understand and have faith in AI-driven results.

3.4 AI ALIGNMENT

The perpetual advancement and evolution of AI systems necessitate the crucial task of ensuring that their objectives align with human values, intents, and ethical norms. The risks are exacerbated by the growing capabilities and deployment in critical domains. When designing an AI system, it is crucial to ensure that its actions and outcomes are consistent with human desired intentions and expectations. In a nutshell, AI alignment is the process of guaranteeing that AI systems adhere to human values and objectives, with a greater emphasis on the objectives of AI systems than their capabilities. As AI systems evolve, ethical concerns emerge when their actions deviate from human expectations, a phenomenon known as the misalignment of AI systems. To maximize AI's benefits and reduce its potential risks, it is imperative to address the misalignment problem Gabriel (2020).

Some misalignment behaviors are less risky but can be dangerous in critical domains, including medical and nuclear fusion control. Moreover, the chances of misalignment increase as the AI system becomes more complex. For example, deep learning-based systems have an increasingly large impact on alignment. Similarly, undesirable behaviors exhibited by large language models (LLMs), such as providing false information, excessive flattery, and deception, become more pronounced as the model size increases. This raises questions regarding the control and management of advanced AI systems, raising concerns regarding the ethical implications. The report of the Safety Summit organized by the UK Government (2024) provides potential risks of AI to individuals and public safety & privacy. The report also emphasizes that the effects of technology on society cannot be adequately handled just through the use of technology. Considering AI alignment solely as a technological issue misplaces authority and control. It is inappropriate for technologists to be the sole decision-makers when determining which risks and values are essential. Social

scientists should collaborate to define and specify all important human values, objectives, and ethical implications. Alignment research and practice address the risks that result from misaligned systems. Hence, the subsequent sections first identify the reasons and risks of misalignment and then present the approaches used to improve the AI alignment.

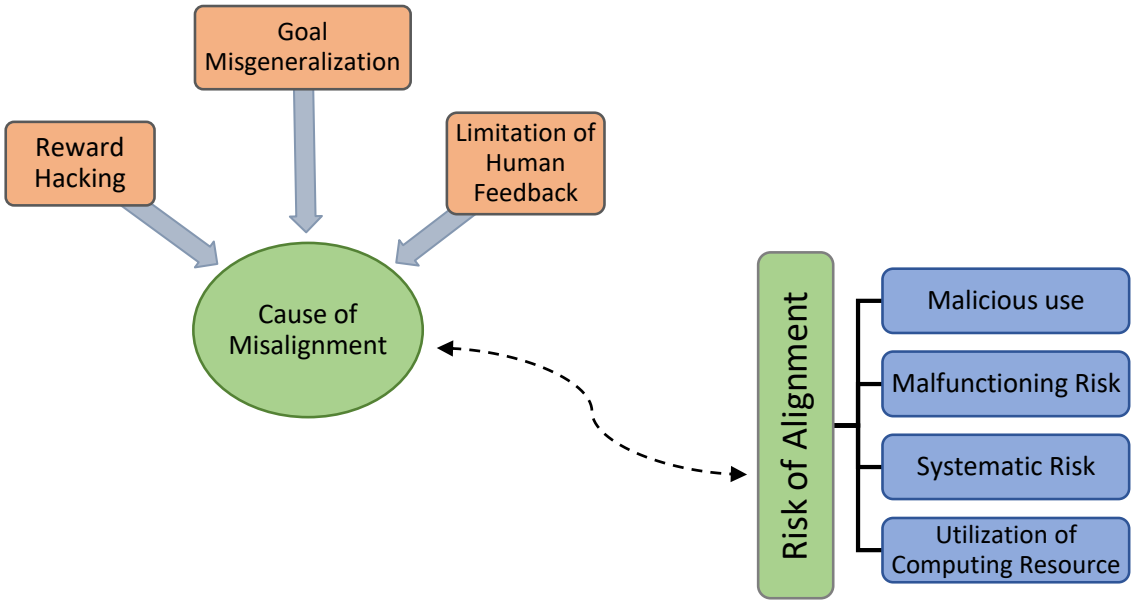

FIGURE 3.4 : AI alignment Risk and Causes of Misalignment

3.5 CONCLUSION

This chapter (is based on my publication Rehman *et al.* (2026a)) highlights the importance of safe AI, particularly in safety-critical systems like healthcare, aviation, and finance, where errors can have severe consequences. This chapter addresses the problem defined in the introduction of the thesis research question 1. Safe AI is defined through essential characteristics : fairness, robustness, reliability, and alignment with human values. The chapter discusses the significance of these factors, their taxonomy, and associated challenges, while also reviewing government measures to ensure AI safety. By prioritizing

these pillars, safe AI aims to build trustworthy systems that users can rely on to operate reliably without risking lives or significant financial loss. Together, these criteria form the foundation of trustworthy AI and the impact of that is reflected in my research in the chapters below.

CHAPITRE IV

FEDCIM : HANDLING DATA HETEROGENEITY IN FL TO IMPROVE FAIRNESS AND ROBUSTNESS

This chapter (is based on my publication Rehman *et al.* (2025a)) presents our approach, "**FedCIM**," that effectively handles non-IID data in collaborative learning, addressing limitations such as predetermined cluster counts, client participation, and inferential datasets. We introduce a clustering approach that dynamically optimizes the cluster while improving fairness and robustness. We showed the efficiency of our approach by achieving an accuracy of 95.33%, via an empirical investigation on diversified benchmark datasets.

4.1 FEDCIM : CLUSTER FL WITH AN ITERATIVE MECHANISM

We considered clients' diverse data distributions as separate tasks and applied multitask learning to provide personalized models for each cluster. The underlying idea is to iteratively minimize the average training loss of cluster models for all clients while improving the estimation of cluster identification.

The overall process of our approach is illustrated in Fig. 4.1. FedCIM utilized a standard FL system comprising a set of honest clients $N = \{1, .., n\}$ and a trusted centralized server. The server is responsible for initializing and controlling the overall FL process. The clients participate in the FL process upon the server's selection and train the global model using their local data. The objective is to learn a global model collaboratively without exchanging the local data by minimizing a global loss function $F(w)$. The global loss function is the weighted average of the local loss function of N participants.

$$minF(w) = \frac{1}{N}\sum_{k=1}^{N} p_k F_k(w) \tag{4.1}$$

Where $F_k(.)$ is the local loss function for a client $k\ \varepsilon\ N$ with a local dataset D_k and $n_k = |D_k|$. Here p_k is the weight of the client k such that $p_k >= 0$ and $\sum_{k=1}^{n} p_k = 1$.

$$F_k(w) = \sum_{j_k=1}^{n_k} f_{j_k}(w; x^{j_k}, y^{j_k}) \tag{4.2}$$

We consider the widely used optimized version of the FL algorithm, Federated Averaging (FedAvg) McMahan *et al.* (2016), based on a simple model averaging aggregation strategy. We utilize FedAvg to average the local cluster models because it allows collaborative learning of a unique global model. The server uses the local updates provided by clients during federated training rather than requesting simultaneous transmission of their parameters. This eliminates the need for additional communication between the server and clientsEspinoza Castellon *et al.* (2022).

4.1.1 TRAINING PROCESS

A classical FL training process executes the following steps until the specific stopping criteria are met. The training process comprises multiple rounds ; each round consists of the following three steps. A typical FL training round is as follows :

1- Client selection and parameter sharing : The server selects a subset of clients. Each selected client downloads the latest global model parameter w_t from the server. For the first iteration $t = 0$, every client has the same initial parameter vector as the server.

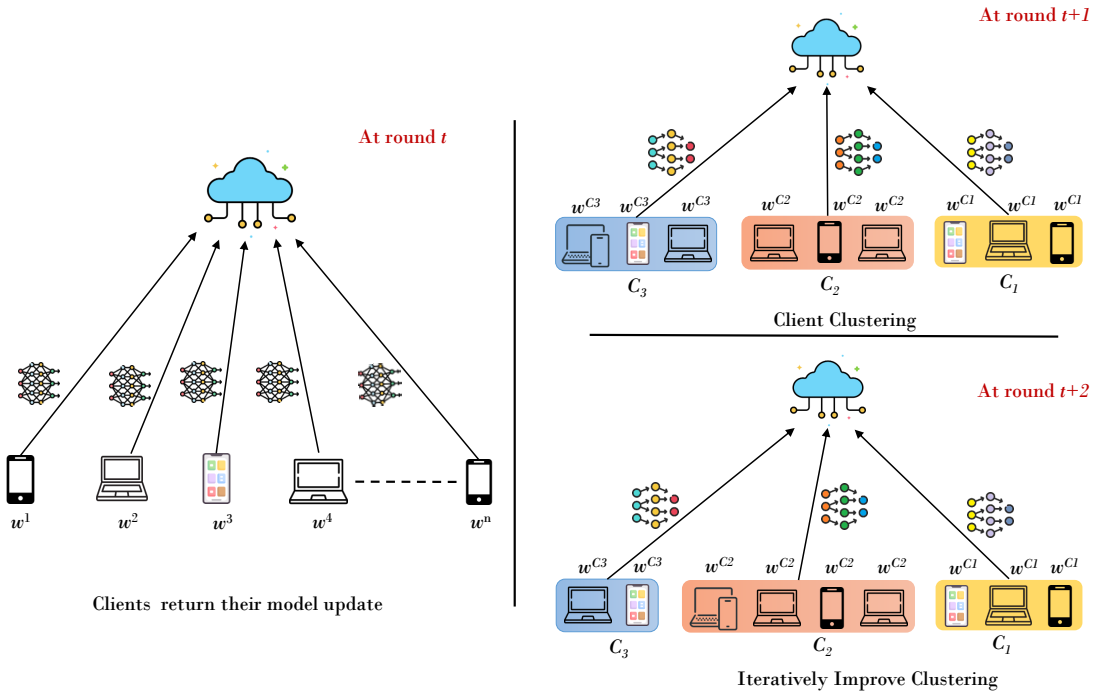

FIGURE 4.1 : An example of the FedCIM Cluster FL process. At round t, all the selected clients return their local model update to the server. The FL server then clusters clients based on the similarity between their return model updates. At round t+2, FedCIM iteratively improves the clients' clustering.

2- Local training : Every selected client performs the local training for each local iteration with its local dataset D_k and shares the gradient update $\Delta_t^{(k)} = w_t - w_t^{(k)}$ with the server.

3- Model Aggregation : The server takes the gradients from the participant clients, aggregates these gradients to get the global model update, $\Delta_t = \sum_{k=N_t} p_k \Delta_t^{(k)}$ and updates the global model as $w_{t+1} = w_t - \eta \Delta_t$, where η is the learning rate of the server.

FedAvg is limited in handling skewed data distributions across clients Zhang *et al.* (2022), potentially causing suboptimal model generalization due to bias toward a client's skewed distribution. However, instead of averaging all the clients' model updates, we cluster the clients with similar data distribution. The clients in a cluster have a homogeneous data distribution, and then aggregate client models update within the cluster. The server then

aggregates all the cluster model updates to obtain the global model. The iterative clustering and the outlier detection process are shown in Algorithm 1. The subsequent sections further illustrate these processes.

4.1.2 ITERATIVELY CLUSTERING PROCESS

Our iterative clustering aims to group clients with a similar data distribution without predefining the number of clusters or requiring the participation of all clients at every round. Hence, we follow the delay clustering strategy for the initial fixed number of rounds. Because the model weights usually remain significantly divergent in the initial rounds, particularly in non-IID contexts. Enhancing the model's stabilization through several iterations of basic FedAvg can assist the global model in approaching a satisfactory condition before using iterative clustering techniques. Moreover, premature clustering of clients with significantly different model updates may provide insignificant clusters.

To measure the similarity between the client model updates, we employed cosine-similarity Han *et al.* (2012), a mathematical metric used to calculate the pairwise similarities between two data points. A higher degree of similarity between the model updates of any two clients suggests a higher level of similarity in their data distributions. In Sattler *et al.* (2019), the author demonstrated that the cosine similarities using client weight updates provided better separation than the gradient update. Hence, in our method, we compute cosine similarities based on weight updates instead of gradients. Utilizing the similarity matrices, we categorize clients to ensure that those within a group have a homogeneous data distribution. The cosine similarity between two clients FL model updates $\Delta\theta_i$ and $\Delta\theta_j$ is :

$$Cosine(\Delta\theta_i, \Delta\theta_j) = \frac{\langle \Delta\theta_i, \Delta\theta_j \rangle}{\|\Delta\theta_i\| \|\Delta\theta_j\|} \tag{4.3}$$

It returns a scalar value range $[-1, 1]$, where a value closer to 1 signifies an increase in similarity and a value closer to -1 indicates greater opposition between the two data points.

Subsequently, we employ k-medoidsPark & Jun (2009) as the clustering methodology to categorize the N clients into distinct clusters. Because K-means clustering iteratively finds k centroids and assigns objects to the nearest centroid, it is sensitive to outliers. In contrast to k-means, k-medoids clustering based on the most centrally located object is less sensitive to outliers and is more efficient in computational time.

To alleviate the need for a predefined number of clusters for the client's clustering, we compute the silhouette valueBelyadi & Haghighat (2021) as in equation 4.4. The silhouette value is defined as the distance between the client i and all other clients within the same cluster, i.e., the $d_{intra}(i)$ means intra-cluster distance, and the distance between the client i and all other clients in different clusters, i.e., $d_{inter}(i)$ the smallest means inter-cluster distance. A higher silhouette value indicates a greater likelihood of the client i being accurately assigned to the appropriate cluster.

We now define the *silhouette* value of one data point i

$$s(i) = \frac{d_{inter}(i) - d_{intra}(i)}{\max\{d_{intra}(i), d_{inter}(i)\}}, \text{ if } |C_I| > 1 \tag{4.4}$$

Where C_I is the cluster from which the data point i belongs. If there is only one cluster, then

$$s(i) = 0, \text{ if } |C_I| = 1$$

Which can also be written as :

$$s(i) = \begin{cases} 1 - d_{intra}(i)/d_{inter}(i), & \text{if } d_{intra}(i) < d_{inter}(i) \\ 0, & \text{if } d_{intra}(i) = d_{inter}(i) \\ d_{inter}(i)/d_{intra}(i) - 1, & \text{if } d_{intra}(i) > d_{inter}(i) \end{cases}$$

From the above definition, the silhouette score ranges within [-1, 1].

$$-1 \leq s(i) \leq 1$$

To reduce the need for all clients to participate in every training round, after the clustering is complete, the server selects random clients from each cluster for the next round. After each training round, the client within the cluster follows the FedAvg algorithm and sends the cluster's global model to the server. The silhouette score iteratively improves the clusters by measuring the distance between the clients of similar clusters and the adjacent clusters. This iterative refinement of clusters allows for dynamic adaptation of the clusters, leading to improved convergence. Upon receiving the cluster global model, the server applies FedAvg to all the cluster models and obtains a global model.

To provide robustness and tackle malicious clients, we employed outlier detection in our algorithm to detect potential outliers during the clustering of clients. It involves measuring the distance between each client's update and the centroid of their assigned cluster. The outlier threshold parameter controls the system's sensitivity to outliers. If the outlier data distribution is closer to another cluster, it is reassigned to that cluster. Only clients that cannot be reassigned to any cluster are considered true outliers and are excluded from the aggregation process. This

Algorithm 1 Iterative Clustering FedCIM

Require: Initial set of clients $\{1,\ldots,n\}$, hyperparameters K_{min}, K_{max}, E, τ
Ensure: Final clustering $\mathscr{C}$ and aggregated models for each cluster

Initialize global model $\mathscr{W}_G$ with CNN architecture
$\mathscr{C}_0 \leftarrow \{1,\ldots,n\}$ {assign all clients to initial cluster}
Models $\leftarrow \{\mathscr{W}_G\}$ {initialize models set with global model}
for each communication round r **do**
  deltas $\leftarrow \{\}$
  for each client $i \in \{1,\ldots,n\}$ **do**
    $\mathscr{W}_i \leftarrow \mathscr{W}_G$ {initialize client model}
    for epoch $e = 1$ to E **do**
      Update $\mathscr{W}_i$ using SGD on local data $\mathscr{D}_i$
    end for
    deltas $\leftarrow$ deltas $\cup\{\mathscr{W}_i - \mathscr{W}_G\}$
  end for
  `// Clustering Phase (Fairness)`
  $S \leftarrow$ ComputeSimilarityMatrix(deltas)
  $D \leftarrow 1 - S$ {convert to distance matrix}
  best_score $\leftarrow -1$
  for $k = K_{min}$ to K_{max} **do**
    $\mathscr{C}_k$, $s_k \leftarrow$ KMedoidsClustering(D,k)
    if $s_k >$ best_score **then**
      $\mathscr{C} \leftarrow \mathscr{C}_k$
      best_score $\leftarrow s_k$
    end if
  end for
  `// Outlier Detection (Robustness)`
  outliers $\leftarrow \{\}$
  for each cluster $\mathscr{C}_j \in \mathscr{C}$ **do**
    for each client $i \in \mathscr{C}_j$ **do**
      $\mu_{ij} \leftarrow$ MeanSimilarity$(i, \mathscr{C}_j, S)$
      if $\mu_{ij} < \tau$ **then**
        outliers $\leftarrow$ outliers $\cup\{i\}$
      end if
    end for
  end for
  `// Model Aggregation`
  cluster_models $\leftarrow \{\}$
  for each cluster $\mathscr{C}_j \in \mathscr{C}$ **do**
    $\mathscr{W}_{\mathscr{C}_j} \leftarrow \frac{1}{|\mathscr{C}_j|}\sum_{i\in\mathscr{C}_j}\mathscr{W}_i$
    cluster_models $\leftarrow$ cluster_models $\cup\{\mathscr{W}_{\mathscr{C}_j}\}$
  end for
  $\mathscr{W}_G \leftarrow \frac{1}{|\mathscr{C}|}\sum_{j\in\mathscr{C}}\mathscr{W}_{\mathscr{C}_j}$
  Evaluate accuracy on the test set
  Record metrics(accuracy, $|\mathscr{C}|$, best_score)
end for
return $\mathscr{C}$, cluster_models

method preserves valuable clients and is flexible and robust. Additionally, it also improves the performance of the FL process.

4.2 EMPIRICAL STUDY

To evince the effectiveness of our proposition, an empirical investigation is conducted using various parameter settings and data distribution scenarios. We evaluated the performance of our approach and compared it with baseline approaches. This section presents the empirical evaluation, results, and a detailed discussion of the findings.

4.2.1 EMPIRICAL EVALUATION SETTING

All empirical evaluations have been conducted on a machine comprised of an Intel(R) Core(TM) i7-9700 CPU and 32GB of RAM. The implementation of our approach is based on PyTorch Paszke *et al.* (2019) ; the baseline methodologies are also implemented using the same library. Both the server and the clients are simulated on the same machine, supported by the observation that the physical location of the server and clients does not influence the performance metrics we evaluate. All the evaluations are executed five times, and we take the average accuracy to mitigate the effect of randomness. The assessment is conducted over 100 communication rounds. We presume that 100 clients are accessible for all evaluations, with 40% randomly selected from each cluster.

We utilized three popular image classification benchmark datasets, including MNISTDeng (2012), FMNISTXiao *et al.* (2017) and CIFAR-10Krizhevsky (2009). These datasets comprise fixed-size images of 70 thousand handwritten digits that are in normalized size, centered, and aligned. We divided these datasets into 60 thousand images for training and allocated the remaining 10 thousand images for testing purposes. Both the training and

testing datasets are assigned to the clients to maintain the proper collaborative learning (FL) setting. The central server receives the model updates from clients and averages them to make the global and cluster models for evaluating each client's local dataset.

We employ a convolutional neural network (CNN) comprising two convolutional layers of 5×5 dimensions, with the first layer featuring 32 channels and the second containing 64 channels. A 2×2 max pooling layer succeeds each convolutional layer, preceded by two fully connected layers that produce a 512-dimensional vector and a 10-dimensional vector, respectively, utilizing an activation function ReLU after each convolutional & fully connected layer and culminating in a softmax output layer to obtain the class probabilities. To evaluate the effectiveness of our techniques, we utilize prominent state-of-the-art methods, FedAvg McMahan *et al.* (2016) and CFL Sattler *et al.* (2019), as baselines for comparison. We have chosen these baselines because they are related to our work, especially CFL, a fast clustering algorithm that recursively bi-partitions the two client groups based on the diversity of the gradient update.

4.2.2 DATA DISTRIBUTION

Although these datasets are intrinsically IID, we generated non-IID scenarios to validate our findings. We simulate the following distinct types of data distribution settings to comprehensively evaluate the performance of our approach.

— **IID :** The training and testing data are randomized and evenly distributed among the number of clients. Each client possesses an equal number of training and testing images with equal target classes. The training data comprises 60K examples, and the testing data contains 10K images.

— **Non-IID Case 1 :** To replicate the data heterogeneity encountered in FL, we partition the training images into two segments per class according to specified ratios (25%, 50%, and 75%), indicating varying levels of non-IID. A single segment is partitioned into a total number of clients, with each component of training data comprising an identical quantity of training images from the same category. The remaining segment is partitioned similarly, with each unit of training data comprising an equivalent number of training photos from various categories. Every client receives two sets of training data from each component. The testing data are also processed identically.

— **Non-IID Case 2 (Single label class) :** We imitate the extreme scenario in FL, wherein each client possesses just one class of label data. The training images are equally partitioned into a total number of clients, with each segment including training images from the same class. The testing images undergo processing identical to that of the training data. Each client is provided with two pieces of training data and testing data from the same class.

— **Adversarial Robustness :** We establish a scenario in which we introduce random (malicious) updates following 70 rounds of training on the FMNIST dataset. The system undergoes training for the first 70 rounds to enable the model to sufficiently learn from other clients before initiating the attack with an adversarial rate of 0.7%. The adversarial rate signifies the proportion of the training dataset that has been modified or compromised by an attacker. It is computed as the ratio of modified samples to the total training samples.

We utilized the same SGD for all the investigations as the local optimizer, with a local batch size of 10 and a local epoch of 10. We used the identical number of clusters specified in the original publications for CFL. For all models, we used a learning rate of 0.1.

However, for FedAvg, the momentum was 0.9. We assess the performance of all approaches by calculating the average final local test accuracy across all clients.

4.3 RESULTS ANALYSIS

4.3.1 FAIRNESS

We present statistics on the distribution of test accuracy for all baseline approaches using the benchmark datasets for all data distribution scenarios (IID, non-IID with ratios 25%, 50%, and 75%, and non-IID with a single label class). These statistics include the average test accuracy, the performance of the best 10% and the worst 10% user models. The average test accuracy itself does not reflect fairness. However, the average test accuracy of the worst 10% user models, which have the lowest performance and variance, further supports the notion of fairness.

IID SITUATION :

Table 4.1 shows that FedAvg performed better in IID settings among all methods. The CFL exhibits better accuracy, slightly worse than FedAvg, because it considers all the clients in one cluster (as shown in Fig. 4.2(a)). This improvement is because CFL utilizes the same aggregation and training methods as FedAvg, with the primary difference being the clustering action. Considering the client as one cluster, it results in performance near FedAvg. Our model performs the least, but is relatively close to FedAvg due to the iterative nature of clustering clients.

NON-IID CASE 1 :

We notice significant variations in performance in non-IID settings. For this case, we run the evaluation for diverse data distribution scenarios. According to Table 4.1, FedAvg exhibits considerable fluctuations in accuracy due to the intrinsic non-IID characteristics of the data distributions between clients. It averages all the clients' local models, which critically affects the accuracy of the global model. CFL, on the other hand, is inconsistent in clustering and did not consider data distribution case 1, as evidenced by Fig 4.2(b). It ran the FL process for many rounds and performed clustering after it converged to a specific point. It acts as FedAvg, putting all the clients in one cluster, which is futile for the purpose of clustering clients. Table 4.1 shows that the performance of CFL (i.e., 92.17%) for case N1 is better than ours. The rationale is that clients in CFL are categorized into clusters that ensure the data is as IID as feasible within each group, significantly enhancing performance. However, our model performs iterative clustering, yielding excellent performance (93.13%) in diverse data distribution environments. As the data distribution increases, the performance of our model increases linearly.

NON-IID SINGLE LABEL CLASS :

In a worst-case scenario of non-IID, each client contains data for only one label class. Table 4.1 illustrates that our technique performs better among all approaches in the worst-case non-IID (single label class) scenario. The reason is that we follow the iterative approach, which focuses on checking the outlier. Outliers are identified by calculating the distance between a client's gradient and the centroid of their assigned cluster. Clients exceeding the outlier threshold are flagged as outliers. If closer, the client is reassigned to another cluster, and true outliers are excluded. This can be seen in Fig. 4.3, which presents

TABLEAU 4.1 : Complete Performance statistics on different data distributions for benchmark datasets(N1, N2, N3 stands for Non-IId Case 1 and N4 shows Non-IID Case 2)

Dataset	Method	Type	IID	N1 (25%)	N2 (50%)	N3 (75%)	N4
FedAvg	MNIST	Average(variance)	**0.988(2.11e-6)**	0.685(6.94e-2)	0.691(6.37e-2)	0.695(6.01e-2)	0.102(6.98e-2)
		Worst 10%	**0.963**	0.346	0.357	0.365	0.003
		Best 10%	**0.995**	0.804	0.816	0.821	0.443
	FMNIST	Average(variance)	**0.914(3.29e-4)**	0.182(7.38e-4)	0.194(7.13e-4)	0.201(7.01e-4)	0.099(9.32e-4)
		Worst 10%	0.891	0.099	0.103	0.011	0.006
		Best 10%	0.960	0.407	0.413	0.419	0.351
	CIFAR	Average(variance)	0.512(6.54e-8)	0.409(5.36e-8)	0.413(5.13e-8)	0.425(5.04e-8)	0.087(7.10e-8)
		Worst 10%	0.204	0.187	0.193	0.199	0.002
		Best 10%	0.763	0.429	0.435	0.446	0.356
CFL	MNIST	Average(variance)	0.947(3.21e-5)	**0.921(7.35e-8)**	0.913(7.30e-8)	0.906(7.25e-5)	0.691(7.96e-2)
		Worst 10%	0.922	**0.915**	**0.889**	**0.881**	0.546
		Best 10%	0.948	0.945	0.944	0.939	0.792
	FMNIST	Average(variance)	0.890(9.65e-5)	**0.859(4.35e-4)**	0.848(5.09e-8)	0.839(5.21e-4)	0.880(7.18e-2)
		Worst 10%	0.885	**0.872**	**0.859**	**0.850**	0.754
		Best 10%	0.912	0.902	0.899	0.894	0.960
	CIFAR	Average(variance)	0.951(7.39e-5)	0.652(5.11e-2)	0.646(5.01e-2)	0.639(5.28e-2)	0.656(6.52e-2)
		Worst 10%	0.910	0.353	0.345	0.339	0.455
		Best 10%	0.985	**0.841**	0.836	0.824	0.843
FedCIM	MNIST	Average(variance)	0.938(1.36e-8)	0.915(4.08e-4)	**0.920(3.99e-8)**	**0.931(3.87e-2)**	**0.953(3.34e-8)**
		Worst 10%	0.920	0.779	0.781	0.788	**0.936**
		Best 10%	0.961	**0.954**	**0.958**	**0.964**	**0.993**
	FMNIST	Average(variance)	0.910(3.12e-6)	0.857(6.09e-10)	**0.864(5.99e-10)**	**0.875(5.72e-10)**	**0.984(2.05e-6)**
		Worst 10%	**0.943**	0.436	0.446	0.451	**0.989**
		Best 10%	**0.999**	**0.933**	**0.940**	**0.949**	**0.993**
	CIFAR	Average(variance)	**0.966(1.64e-8)**	**0.681(6.86e-3)**	**0.689(6.14e-3)**	**0.696(5.11e-2)**	**0.992(2.63e-6)**
		Worst 10%	**0.987**	**0.363**	**0.379**	**0.382**	**0.992**
		Best 10%	**1.0**	**0.838**	**0.849**	**0.858**	**0.996**

performance per round, silhouette score, and number of clusters per round. We noticed that FedAvg performs poorly when facing severe cases of non-IID. Conversely, CFL shows improved performance, worse than the FedCIM approach. Fig. 4.2(c) indicates that CFL runs a normal FL process for some initial rounds and then starts clustering based on client similarity and bi-partitioning criteria until a specific condition is met. However, we can see in the figure that CFL clustered one client into multiple clusters and did not consider some clients in the clustering process. In this CFL setting, we set the number of clients to 10 to show the results on the graph. The result raises a serious concern for the performance and efficiency of CFL.

The results indicated that our iterative clustering method improves performance under non-IID conditions. However, under extreme non-IID situations, the iterative process, in conjunction with an outlier detection process, yields significantly improved results.

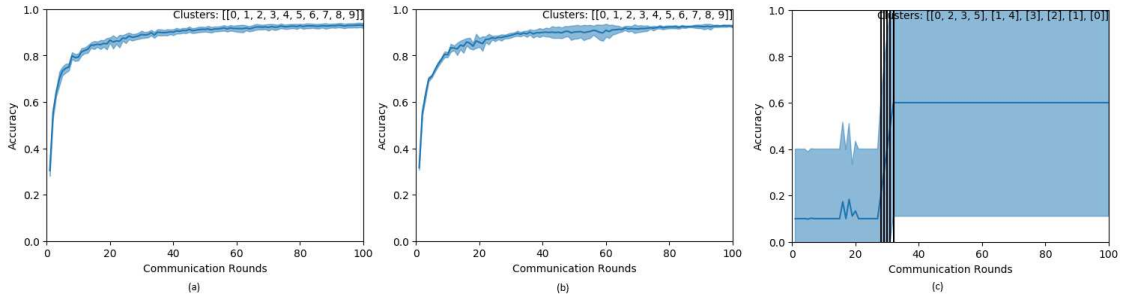

FIGURE 4.2 : CFL performance in all cases. Fig.(a) shows the performance in the iid setting ; (b) shows the performance in the non-iid case 1 (75%) ; (c) shows the non-iid case 2.

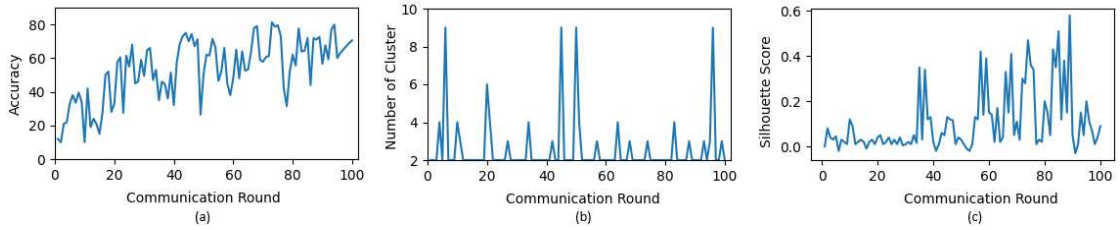

FIGURE 4.3 : Performance of FedCIM in Case 2. Fig.(a) shows the fairness accuracy ; Fig.(b) shows the number of clusters per communication round, which changes according to the change in the silhouette score (c)

4.3.2 ROBUSTNESS

This section presents the effectiveness of our method in addressing adversarial robustness (adversarial and Byzantine attacks) through the use of the FMNIST dataset. To illustrate the effect of adversarial clients, we established the same scenario outlined in the preceding section. We assigned sixty percent of clients as adversaries to deliberately alter their local model updates, consequently introducing misleading model updates into

the global aggregation process. Our approach mitigates the impact of adversarial clients, ensuring that they cannot influence the process. Fig. 4.4 (a) illustrates that our approach successfully eliminates adversarial (outlier) clients and then continues the FL process. Then, we perform the same evaluation without eliminating adversarial clients as in Fig. 4.4(b) to exhibit the robustness (50%) of our approach. The global model improved steadily between rounds 2 and 70, quickly converging to high accuracy before the attack. The poisoning attack resulted in a significant decline in accuracy, revealing the model's vulnerability to malicious updates. To compare the performance of our approach with CFL, we run it with the same settings as ours; CFL does not provide robustness against adversarial clients. Fig. 4.5 shows CFL's performance with and without adversarial clients. We ran the normal process until the 49th communication round, and at the 50th round, we started the attack. The result indicates that the CFL cannot handle adversarial clients, and its accuracy decreased drastically after the attack. This shows that the conjunction of our iterative clustering and outlier detection process helps in improving the robustness of our approach.

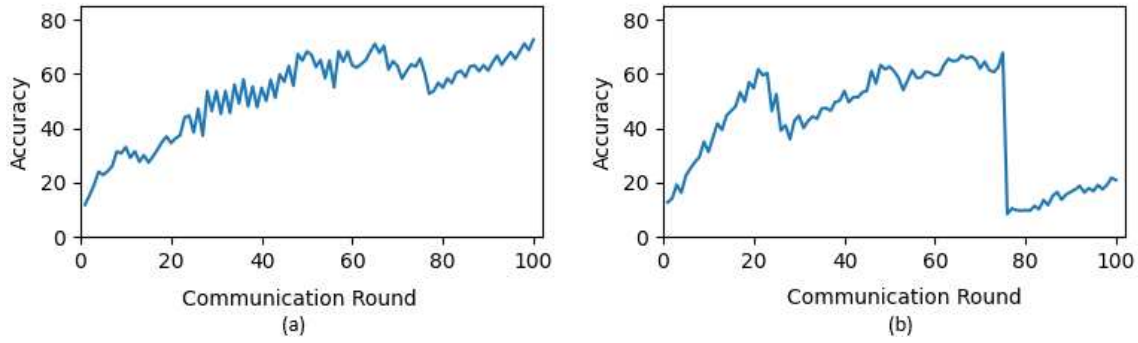

FIGURE 4.4 : Robustness of FedCIM in detecting the byzantine attack and outlier clients

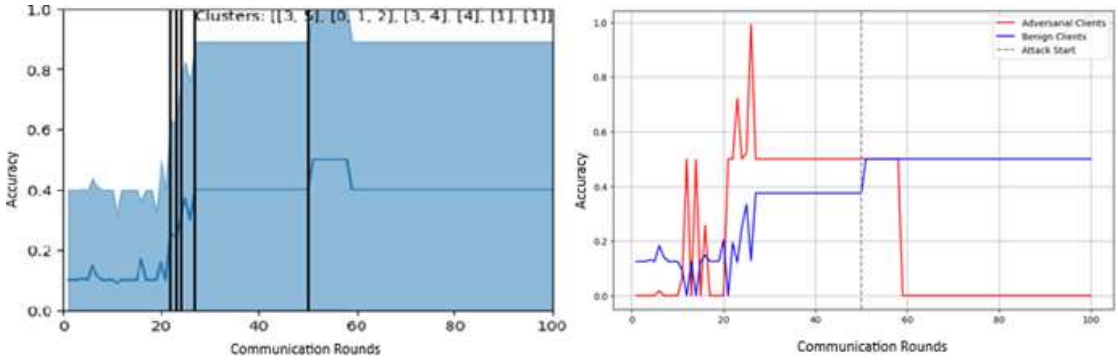

FIGURE 4.5 : Robustness of CFL for adversarial clients

4.4 DISCUSSION

This section discusses the important facts we observed during the empirical study.

4.4.1 INFLUENCE OF LOCAL EPOCHS

The number of local epochs influences the performance of global or personalized cluster models. For instance, as shown in Table 4.2, our investigation depicts the increase in performance with a rise in the number of local epochs. The table illustrates the correlation between the performance of our model and the number of local epochs with three different scenarios. A single local epoch results in fewer local updates from clients, resulting in slow training and less accuracy than scenarios with multiple local epochs, assuming constant communication rounds. Additionally, clients who have not received enough training may produce erroneous inference results on the server side, which might impact clustering quality. Conversely, as illustrated in Table 4.2, an excessive number of local epochs, i.e., 20, results in a decline in accuracy (88.97%) due to a significant drift between the server-side averaged model and the clients' local models.

TABLEAU 4.2 : Influence of Number of Local Epochs on Performance (N1, N2 stands for Non-IID Case 1 & Case 2)

Epoches	N1 (25%)	N1 (50%)	N1 (75%)	N2 (Single Class)
1	83.10	83.61	84.17	88.28
10	92.90	93.10	93.40	95.34
20	85.54	86.10	86.64	88.97

4.4.2 TARGET APPLICATION AND DOMAIN RELEVANCY

We have proficiently employed MNIST, Fashion-MNIST, and CIFAR-10 as benchmark datasets, which are conventional in ML research. These datasets offer realistic contexts for the validation of our assessment. Our approach addresses scenarios when clients exhibit diverse data distributions (e.g., geographical disparities in healthcare imaging or personalized device usage patterns), such as :

— Healthcare : Federated training across hospitals with non-IID patient data (e.g., MNIST /FMNIST for medical imaging, CIFAR-10 for histopathology).

— Smart Devices : Personalized recommendations on edge devices (e.g., FMNIST for wearable sensor data).

— Autonomous Systems : Collaborative learning across vehicles with localized data (CIFAR-10 for road sign recognition).

MNIST and FMNIST simulate low-complexity, label-skewed settings, whereas CIFAR-10 evaluates scalability to more complex feature spaces. Future work will extend to domain-specific datasets like CheXpert and human activity monitoring.

4.4.3 THREATS TO VALIDITY

This section examines the possible threats to the validity of the evaluation conducted in this study through adherence to a specified set of criteria Wohlin *et al.* (2012). We

investigated construct, internal, reliability, and external validity to guarantee the robustness and generalizability of our findings.

Threats to Construct Validity, concern the correlation between theory and observations. Our study utilizes three benchmark datasets to replicate non-IID and imbalanced data distributions. Although this dataset is used extensively in FL research, it may not adequately represent the richness and diversity of real-world data distributions. The generating procedure of non-IID data may include biases that inadequately represent the diversity of real-world data distributions. A possible concern is the accuracy and reliability of the clustering algorithm. The metrics that assess the similarity of clients' data distributions may influence the algorithm's efficacy. If the metrics fail to capture the true data properties accurately, it could lead to incorrect clustering, impacting overall efficiency and accuracy. The clustering approach may generate errors if it is not well-tuned to address the differing levels of data imbalance across clients.

Threats to Internal Validity, Several factors may affect the internal validity of our findings. The investigation was performed in a controlled laboratory setting with predetermined client behavior. This may not accurately represent real-world FL scenarios with dynamic client participation and varying data distributions over time. We performed evaluations with various random seeds to mitigate variability and reported the average results. The hyperparameters of our empirical study (e.g., similarity threshold, learning rate, number of communication rounds) may substantially influence the outcomes, and we mitigated this by performing hyperparameter tuning. Additionally, communication latency and bandwidth in FL systems can impact performance ; we did not investigate all potential network conditions. However, iterative clustering affects the computational complexity of our method. Although our clustering enhances model personalization, it incurs a cost of $O\left(k(n-k)^2\right)$ every iteration. For $n = 100$ and $k = 5$, this incurs approximately 1.2

seconds each round on the MNIST dataset. We did not consider the time and computational complexity in this article. To address this, we recommended two strategies. First, substitute k-medoids with more expedient alternatives such as MiniBatch k-means $O(nkd)$, accepting a slight compromise in accuracy for enhanced speed. Secondly, we can diminish the frequency of clustering (e.g., per $R = 5$ rounds) since client distributions evolve slowly. Finally, the efficiency of client clustering may be affected by variables such as model architecture or optimizer setups, which can vary across various applications.

Reliability Validity Threats, pertains to the feasibility of duplicating this study. To guarantee reproducibility, we provided all essential information on our empirical evaluation, encompassing dataset description, hyperparameter settings, and assessment processes. Furthermore, we will publicly release our source code and implementation of the adaptive client clustering methodology. This encompasses the FL framework, data partitioning scripts, and clustering algorithms employed in our research. All data utilized in this empirical investigation are accessible online [1].

Threats to External Validity, concern the generalizability of our findings to real-world scenarios. While our investigation demonstrates the effectiveness of client clustering in simulated non-IID and imbalanced settings, the results may not generalize to all FL applications. For instance, focusing on a specific set of clients and datasets may not fully represent the heterogeneity of data distributions in practical deployments, such as healthcare or finance. Furthermore, our investigation assumes a controlled environment with reliable client and central server communication. However, clients' behavior in FL systems, such as device heterogeneity, network conditions, and device failures, may vary in real-world applications, potentially limiting the generalizability of our findings to specific contexts. The method's scalability to larger clients or complex data distributions is not fully explored,

1. https ://github.com/FedCIM/FedCIM

and additional constraints, such as privacy-preserving mechanisms, could influence the effectiveness and efficiency of the proposed method.

4.5 CONCLUSION

We presented an efficient approach (my publication Rehman *et al.* (2025a)) to improve the fairness and robustness of federated learning models by handling data heterogeneity. This chapter addresses the problem defined in the introduction of the thesis research question 2. We proposed the use of iterative clustering, which iteratively enhances the clustering process. We evaluated our approach empirically and showed how it improved the fairness and robustness of federated learning models. Concretely, our technique outperforms FedAvg and CFL under various evaluative conditions. For iterative clustering, the time and computational cost are higher than those of others. Conversely, minimizing resource consumption in clustering will be essential for future improvements.

CHAPITRE V

RBFL : SECURING FL AGAINST DATA POISONING USING A REPUTATION-BASED APPROACH

This chapter (is based on my publication Rehman *et al.* (2025b)) introduced a fair contribution evaluation approach that employs Cosine-Gradient to evaluate the coalition and TMCShapley to fairly estimate each client's marginal contribution. We utilize a reputation-aware mechanism to guarantee the credibility of the client's input and collaboration robustness against malicious or free-rider clients. Our empirical examination demonstrates the efficacy of our methodology in the healthcare domain and across several benchmark datasets, attaining robustness with an accuracy of 92% while preserving accuracy.

5.1 METHODOLOGY

We presented an efficient reputation-based approach (RBFL) for securing the FL process by evaluating the client's contribution in training the global model. Our proposed approach employs two functions for coalition quality estimation and each client's marginal contribution. It provides adversarial robustness against poisoning and free riders.

5.1.1 PROBLEM SETTINGS

We consider a standard FL setting with N numbers of malicious clients with local data to train a global model. An honest central server having a reputation history of clients that aggregates the client's local model update to produce a global model. Our approach aims to estimate the quality of the coalition $S \subseteq N$ between clients. Based on the quality of this grand coalition, fairly evaluate the contribution ϕ_i for each client $i \in N$ to training a

global model. Additionally, the center server maintains a reputation-aware mechanism to mitigate the problem of malicious clients and free riders. This ensures that clients with low reputations are filtered out to maintain the model's integrity. Thus, for two clients $i, j \in N$, if $\phi_i(v) > \phi_j(v)$, then client i contributed more to utility, v then client j. The next subsections exhibit (1) a simple training process in a typical FL system. (2) How to assess the quality of a grand coalition. (3) Based on the quality of the grand coalition, how can we estimate the fair contribution of each client ? (4) How reputation awareness maintains the history of client updates and filters out malicious clients. We have presented an overview of these in Fig. 5.1.

5.1.2 TRAINING PROCESS

We utilize the commonly used FL-optimized variant Federated Averaging (FedAvg) McMahan *et al.* (2016), which generates a global model by averaging all local model weights received from the participants' clients based on the number of samples used for training. We utilize FedAvg to average the local model updates because it allows collaborative learning of a unique global model. A classical FL training process executes the following steps until the specific stopping criteria have converged. The training process comprises multiple rounds ; each round consists of the following three steps. A typical round for a coalition $S \subseteq N$ is as follows.

1- Client selection and parameter sharing : The server selects a subset of clients. Each selected client downloads the latest global model parameter w_t from the server. For the first iteration $t = 0$, every client has the same initial parameter vector as the server.

2- Local training : Every selected client performs the local training for each local iteration with its local dataset D_k and shares the gradient update $\Delta_t^{(k)} = w_t - w_t^{(k)}$ with the server.

3- Model Aggregation : The server takes the gradients from the participant clients, aggregates these gradients to get the global model update, $\Delta_t = \sum_{k=N_t} p_k \Delta_t^{(k)}$, and updates the global model as $w_{t+1} = w_t - \eta \Delta_t$, where η is the learning rate of the server.

Upon aggregation, the server calculates the utility of this coalition and estimates the individual contribution based on that utility. Additionally, the server maintains a client's reputation history based on the gradient update received from the clients.

5.1.3 COSINE GRADIENT DATA UTILITY

To evaluate the utility of a coalition of clients, the server employed a cosine gradient to measure the alignment between the gradients of two models in the gradient aggregation phase. In each FL iteration, the gradient vectors represent how the model changes throughout this round of training. In this context, the utility of a coalition between clients is evaluated by computing the cosine similarity (from Equation (4.3)) between the gradient of the global model ∇w_t and the aggregated parameter update $\sum_{i\ \varepsilon\ S} \nabla w_i^t$ of the coalition. The utility of the coalition S is :

$$v(S) = CosineGradient(\nabla w_t, \sum_{i\ \varepsilon\ S} \nabla w_i^t) \tag{5.1}$$

A higher degree of similarity between the global model and the gradient update of the model trained on coalition S suggests a higher level of data usefulness for S. Later, this utility function facilitates the valuation function to estimate each client's marginal contribution.

5.1.4 CONTRIBUTION ESTIMATION

In assessing a participant's contribution inside the FL paradigm, direct access to their local data is strictly forbidden. Consequently, the contribution assessment process is conducted without analyzing local data. SV solely necessitates that the final utility attained by the coalition be computed. This enables it to be a practical notion for fairly evaluating contributions in FL. For each coalition, SV provides the marginal contribution of each client in a coalition. SV ϕ_i for a client i is defined as the expected marginal contribution of the client i when joining any coalition $\mathscr{S}$:

$$\phi_i := \frac{1}{N!} \sum_{\pi \in \Pi_{\mathscr{N}}} [\nu(\mathscr{S}_{\pi,i} \cup \{i\}) - \nu(\mathscr{S}_{\pi,i})]$$

where $\Pi_{\mathscr{N}}$ is the set of all permutations of clients, and $\mathscr{S}_{\pi,i}$ is the coalition of clients preceding the client i in permutation π.

The SV for each client i in FL can be formulated as the weighted average of their marginal contributions in all possible coalitions of clients. Since the exact computation of SV is computationally expensive when we have a large number of clients, we approximate them using TMC Shapley.

TMC SHAPLEY APPROXIMATION

The TMC-Shapley method combines Monte Carlo permutation sampling with an early stopping rule to approximate SV efficiently. Instead of evaluating all $N!$ permutations, sample a subset of permutations and compute the average marginal contribution over these samples :

$$\hat{\phi}_i := \frac{1}{\bar{M}} \sum_{\pi=1}^{M} [\nu(\mathscr{S}_{\pi,i} \cup \{i\}) - \nu(\mathscr{S}_{\pi,i})]$$

where $\bar{M}$ is the number of sampled permutations, and $\mathcal{S}_{\pi,i}$ is the coalition of clients preceding client i in the π-th sampled permutation.

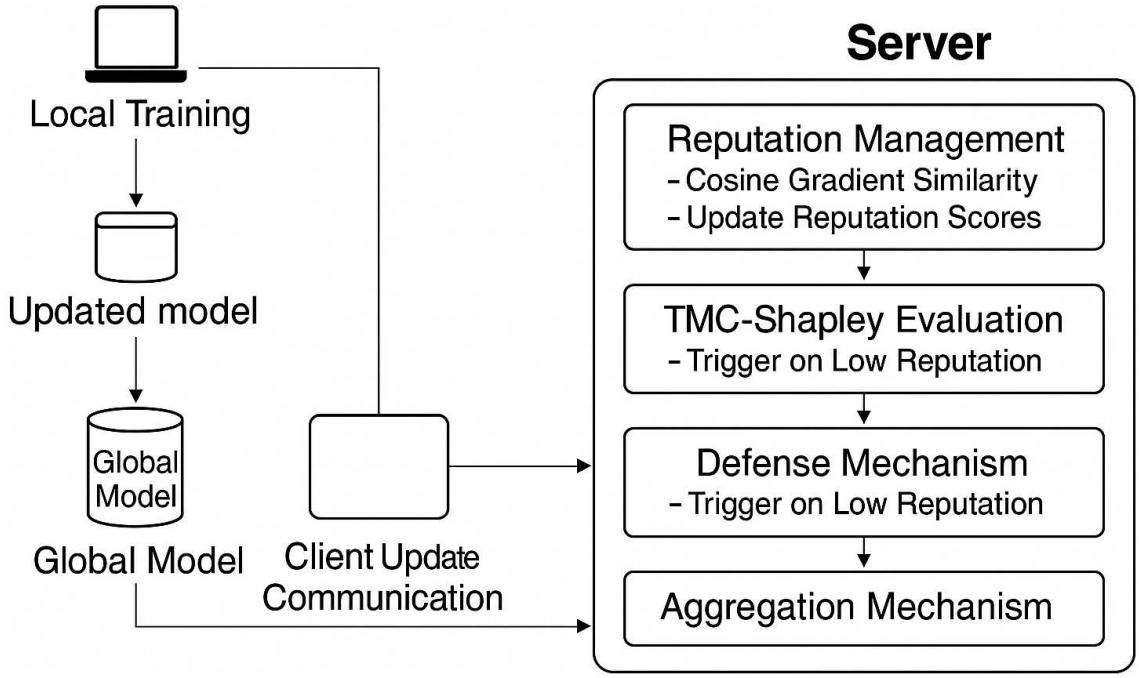

FIGURE 5.1 : Overview of the RBFL : reputation calculation, contribution evaluation, reactive defence, and aggregation.

It set a threshold for early stopping. This represents the performance tolerance where marginal contributions are deemed negligible. If the truncation threshold is met and performance is stabilized (i.e., close to the full dataset performance), truncate the permutation when the marginal contribution falls below τ. This assumes that adding more data points beyond this point has a negligible impact, reducing the number of model retrainings.

Algorithm 2 Reputation-Based Filtering and Defence

Input : Dataset $\mathscr{D}$; number of clients N; rounds R; epochs E; learning rate η; reputation threshold τ; poison rate ρ; attack start round r_a; number permutations m
Output : Final global model $\mathscr{M}_G$; Shapley values Φ
Initialize $R_i \leftarrow 1.0$ for all $i \in \{1,...,N\}$; set $\mathscr{E} \leftarrow \emptyset$
Split dataset $\mathscr{D}$ among N clients (iid / non-iid / clustered)
Select $\lfloor \rho \cdot N \rfloor$ malicious clients : $\mathscr{A} \subseteq \{1,...,N\}$
Initialize global model $\mathscr{M}_G$
for round $r = 1$ to R **do**
    $\mathscr{P}_r \leftarrow \{i \mid i \notin \mathscr{E}\}$
    for each $i \in \mathscr{P}_r$ **do**
        Set model $\mathscr{M}_i \leftarrow \mathscr{M}_G$
        if $r \geq r_a$ and $i \in \mathscr{A}$ **then**
            Go to Label Poisoning Phase
        else
            Train $\mathscr{M}_i$ on clean D_i for E epochs
        end if
        Compute gradient $G_i \leftarrow \mathscr{M}_i - \mathscr{M}_G$
        Evaluate local accuracy A_i
    end for
`// Label Poisoning Attack Phase (only for` $i \in \mathscr{A}$`)`
for each $i \in \mathscr{P}_r \cap \mathscr{A}$ **do**
    Flip labels in a ρ-fraction of D_i randomly
    Re-train $\mathscr{M}_i$ on poisoned D_i for E epochs
    Re-compute $G_i \leftarrow \mathscr{M}_i - \mathscr{M}_G$
    Re-evaluate local accuracy A_i
end for
Aggregate $\{\mathscr{M}_i\}$ using FedAvg $\Rightarrow \mathscr{M}_G$
Compute cosine similarity matrix S_{ij} between $\{G_i\}$
`// Reputation Update Phase`
for each $i \in \mathscr{P}_r$ **do**
    $R_i \leftarrow \alpha R_i + (1-\alpha) \cdot \frac{\text{mean}(S_i)}{\text{mean}(S)}$
end for
`// Reactive Filtering Phase`
if $r \geq r_a$ and reactive defense enabled **then**
    for each $i \in \mathscr{P}_r$ **do**
        if $i \in \mathscr{A}$ and $|\mathscr{P}_r| - |\mathscr{A}| \geq 4$ **then**
            $\mathscr{E} \leftarrow \mathscr{E} \cup \{i\}$
        end if
    end for
end if
`// Shapley Valuation Phase`
Estimate ϕ_i for all i using TMC-Shapley with m permutations
Record : global accuracy, reputations $\{R_i\}$, local accuracies $\{A_i\}$, Shapley values $\{\phi_i\}$
end for
return $\mathscr{M}_G$, Φ

5.1.5 REPUTATION CALCULATION

To introduce reputation awareness, after each round of training, compute the cosine similarity between each client's gradient and the global gradient. Based on the direction of the gradient of the client updates, the reputation score is assigned. Reputation scores are maintained throughout the training rounds and updated based on contribution values. We define a threshold τ to filter out clients with low reputation scores. The client is filtered out in subsequent rounds if the reputation score is below the threshold τ. Filtering out low-reputation clients improve the global model accuracy, demonstrating the integrity of the filtering mechanism and the ability to protect the global model from low-quality updates. Clients with high reputations consistently contribute positively, while the free-riders and malicious clients are filtered out. Algorithm 2 shows the reputation filtering mechanism securing the FL process from poisoning attacks.

5.2 EMPIRICAL STUDY

To evince the effectiveness of our proposition, an empirical investigation is conducted using various parameter settings and data distribution scenarios. We evaluated the performance of our approach and compared it with baseline approaches. This section presents the empirical evaluation, results, and a detailed discussion of the findings.

5.2.1 EMPIRICAL EVALUATION SETTING

All empirical evaluations have been performed on a machine comprised of an Intel(R) Core(TM) i7-9700 CPU and 32GB of RAM. Both the server and the clients are simulated on the same machine, supported by the observation that the physical location of the server and clients does not influence the performance metrics. All evaluations are executed five times,

and we take the average accuracy to mitigate the effect of randomness. The assessment is conducted over 100 communication rounds. We presume that 100 clients are accessible for all evaluations, with 50% randomly selected.

We used well-known benchmark datasets, including MNIST Deng (2012), Fetal-Health de campos *et al.* (2000), and adult Becker & Kohavi (1996) datasets. The popular MNIST image classification benchmark comprises fixed-size images of 70 thousand handwritten digits in normalized size, centered, and aligned. Divide the dataset into 60 thousand images for training and allocate the remaining 10 thousand digits' images for testing purposes. The training and testing datasets are assigned to clients to maintain a proper collaborative learning setting. We employ a simple neural network using the Keras Sequential API, designed for flexibility with varying input shapes and class numbers. The model begins with a flattened layer that transforms the input data, specified by the input shape parameter, into a one-dimensional vector. This is followed by two fully connected layers, 128 and 64 units, employing the ReLU activation function. The final output layer has 10 units, utilizing a softmax activation function to generate class probabilities for a specific classification task.

To evaluate the effectiveness of our approach, we compare our approach with the prominent state-of-the-art methods, FedAvg McMahan *et al.* (2016), Leave-one-outKearns & Ron (1999), and Shapley ValueWang *et al.* (2020). We have chosen these baselines because they are related to our work.

5.2.2 DATA DISTRIBUTION

Although all the utilized datasets are intrinsically IID, we generated non-IID scenarios to validate our findings. We simulate the following distinct types of data distribution settings

to evaluate the performance of our approach comprehensively. Here, we present the example of the MNIST dataset. However, other datasets follow the same distribution settings.

— **(i) S1-IID :** The training and testing data are randomized and distributed evenly among clients, each having an equal number of training and testing images with equal target classes.

— **(ii) S2-Non-IID :** To replicate FL data heterogeneity, training images are divided into two segments per class, each consisting of identical training images from the same category. The remaining segment is divided similarly, with each unit of training data containing an equal number of training photos from different categories. Clients receive two sets of training data from each component, and testing data are processed identically.

— **(iii) S3-Single Label** : We replicate FL's extreme scenario, where each client has one class of label data, with training images divided into clients and testing images processed identically, providing two pieces of data from the same class.

— **Poisoning Attack :** To perform a poisoning attack, we create a scenario in which we inject the label poisoning attack after 70 rounds of training using the MNIST dataset. The system is trained for the initial 70 rounds so that the model learns enough from other clients and then launches the attack with a poison rate of 0.7%. The poisoning rate denotes the fraction of the training dataset altered or compromised by an attacker. It is calculated as the ratio of poisoned samples to the total training samples.

5.3 RESULT ANALYSIS

5.3.1 PREDICTIVE PERFORMANCE

Table 5.1 summarizes the average test accuracy of the global model trained by each client's local data obtained on benchmark datasets against three different data distribution

settings. This table highlights the significant influence of the dataset's characteristics on performance.

In IID distribution (S1), FedAvg performed better than other methods in all datasets (i.e., 97.17%, 91.21%, 92.30% respectively). This is because it averages the model update from all the clients based on the data contributed to their model training. For MNIST, our Shapley variant shows better performance after FedAvg. LOO and SV demonstrate better performance, followed by FedAvg for other datasets.

In the non-IID (S2) setting, we noticed significant variations in performance. According to Table 5.1, FedAvg exhibits considerable fluctuations in accuracy due to the intrinsic non-IID characteristics of the data distributions between clients. It averages all the clients' local models, which critically affects the accuracy of the global model. On the other hand, our approach achieved better accuracy (91.97% & 85.88%) for the MNIST and Adult datasets. This is because the utility function and reputation mechanism up-weight the clients who contributed more to the model training with high similarity to the global model. However, SV shows better accuracy in the FetalHealth dataset (i.e., 91.55%). Our model performs better with an increased dataset size than with a small dataset because TMC-Shapley requires excessively high Monte Carlo iterations to produce reliable estimates.

In a worst-case scenario (S3), each client contains data for only one label class. Table 5.1 illustrates that our approach surpasses alternative methods significantly, especially when dealing with more heterogeneous data partitions. This may be attributed to reputation-weighted aggregation, which can dynamically increase the weight of the clients who contribute more, suggesting that they possess better local data. For the Fetalhealth dataset, SV surpasses our approach with an accuracy of 92.25% because we have noticed that our approach outperforms others with large datasets.

TABLEAU 5.1 : Performance Comparison of RBFL with the Baseline (Accuracy)

Methods	MNIST			ADULT			FETAL_HEALTH		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
FedAvg	**97.17%**	72.94%	70%	**91.21%**	72.39%	73%	**92.30%**	79.73%	78.30%
LOO	97%	90%	88%	86.01%	85.21%	85.24%	89.91%	89.20%	89.90%
SV	95%	89%	90%	80%	77%	71%	92.09%	**91.55%**	**92.25%**
RBFL(SV)	97.11%	70.06%	96.77%	85.43%	85.38%	85.40%	88.50%	84.52%	89.20%
RBFL(TMC)	95.30%	**91.97%**	**96.81%**	85.72%	**85.88%**	**85.84%**	87.80%	88.17%	89.26%

5.3.2 ADVERSARIAL ROBUSTNESS

This section presents the effectiveness of our approach in handling adversarial robustness by utilizing the healthcare dataset (FetalHealthde campos *et al.* (2000)). To demonstrate the impact of adversarial clients, we set up the same scenario as described in the previous section for both non-IID S2&S3 data distributions. We assigned sixty percent of clients as adversaries to intentionally flip a fraction of their local data labels to inject misleading model updates into the global aggregation process. Fig. 5.4 (a) & Fig. 5.5 (a) shows the performance when our approach filtered and eliminated adversarial clients by maintaining a reputation mechanism to reduce the effects of malicious client updates. Then we perform the same evaluation without eliminating adversarial clients as in Fig. 5.4(b) & Fig. 5.5 (b) to exhibit the robustness of our approach. Before the attack, the global model rapidly converged to high accuracy, steadily improving between rounds 2 and 70. However, the poisoning attack led to a massive drop in accuracy, indicating the model's vulnerability to malicious updates. Fig. 5.4 (b) shows the performance before and after recovering from the poisoning attack for non-IID S2. Our approach recovered from the effect of the poisoning attack, with global accuracy reaching around 70% at round 100. However, our solution achieves accuracy near the pre-attack levels, which is around 88%. This indicates that the system was unable to recover from the poisoning attack fully, and the negative impact persists despite gradual recovery. For the non-IID S3, our approach exhibits a much

improved performance of 90% accuracy, which is quite close to the accuracy before the attack starts. The application of reputation filtering improves model accuracy, enhancing stability and robustness. This process reduces the influence of malicious clients, making the system more resilient even in the presence of data poisoning attacks. We compare the performance of our approach with Khuu *et al.* (2024), which employed explainable AI (SHAP) to detect poisoned clients. We chose this approach for our comparative evaluation because it's quite close to our approach and was recently proposed. They compute SHAP values for each client's model update, allowing the server to distinguish anomalous updates by comparing client explanation vectors. The issue with this approach is that the server needed a reference dataset to compare the client's model updates. Fig. 5.2 & 5.3 shows the performance of Khuu *et al.* (2024) with and without handling the poisoning attack for both data distribution cases. We ran the experiments with the same settings as ours and compared the results. Fig. 5.2 indicates that the SHAP-based approach performs better in non-IID S2 than ours. However, in the worst case S3, where every client has only one label class, the SHAP-based approach could not handle adversarial clients, and the accuracy decreased drastically after the attack, as in Fig. 5.3. This approach uses a one-shot explainer to detect an attack after just one affected round. This is because the defense mechanism uses an SVM classifier on the SHAP vectors to separate poisoned updates from normal ones. This means the server performs a client-level inspection of model contributions each round, which is computationally expensive. However, our approach does not catch malicious updates immediately ; the global model may suffer temporary degradation before recovery. Hence, our approach shows much improved performance for the extreme case of data heterogeneity.

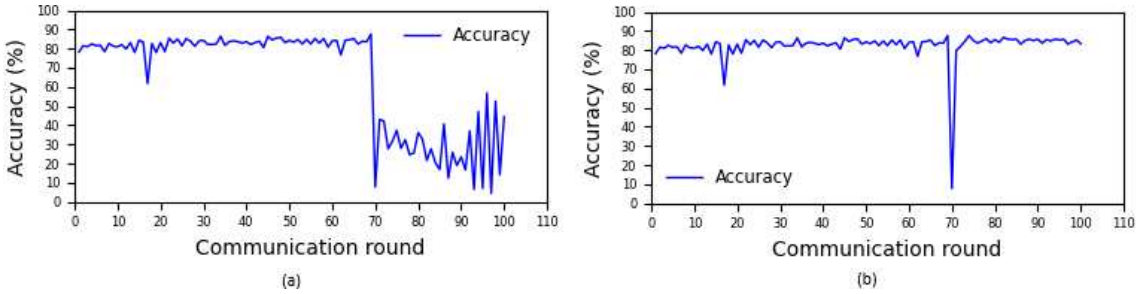

FIGURE 5.2 : SHAP adversarial robustness for non-IID S2 (FetalHealth Dataset).

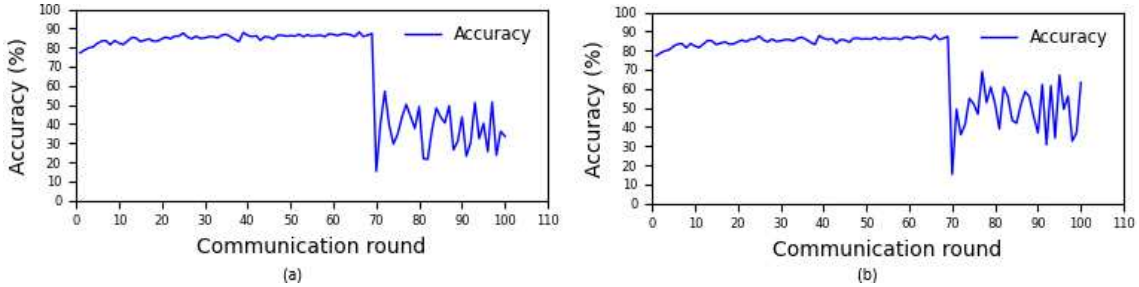

FIGURE 5.3 : SHAP adversarial robustness for non-IID S3 (FetalHealth Dataset).

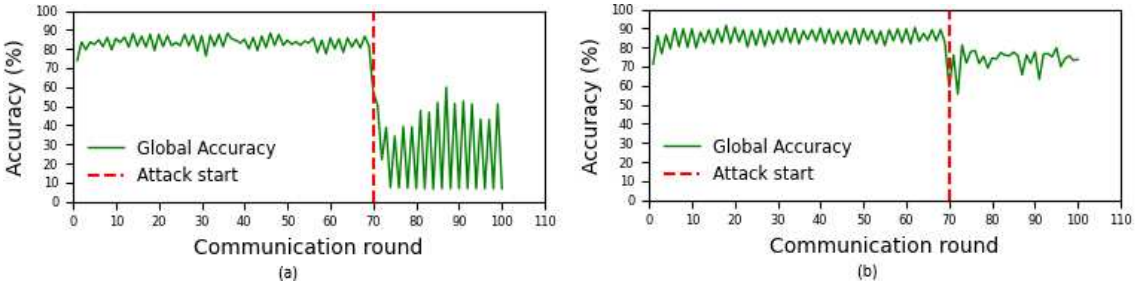

FIGURE 5.4 : RBFL adversarial robustness for non-IID S2 (FetalHealth Dataset).

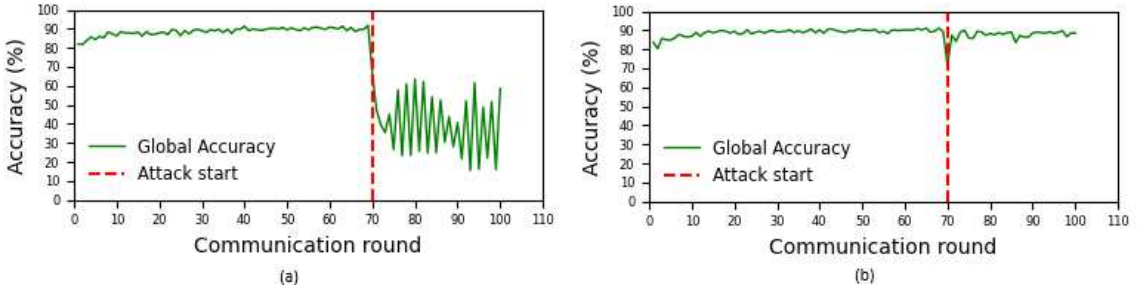

FIGURE 5.5 : RBFL adversarial robustness for non-IID S3 (FetalHealth Dataset).

5.4 DISCUSSION

This section discusses the main facts we observed during the empirical study.

5.4.1 APPROXIMATION OF SHAPLEY VALUE

Since the calculation of SV is computationally expensive when we have a large number of clients, we utilize its approximation, i.e., TMC Shapley. We compare the performance of the SV with its TMC-Shapley approximation on the Adult dataset. Fig. 5.7 shows that TMC-Shapley approximates the contribution values well and achieves a similar or slightly better global model accuracy than SV.

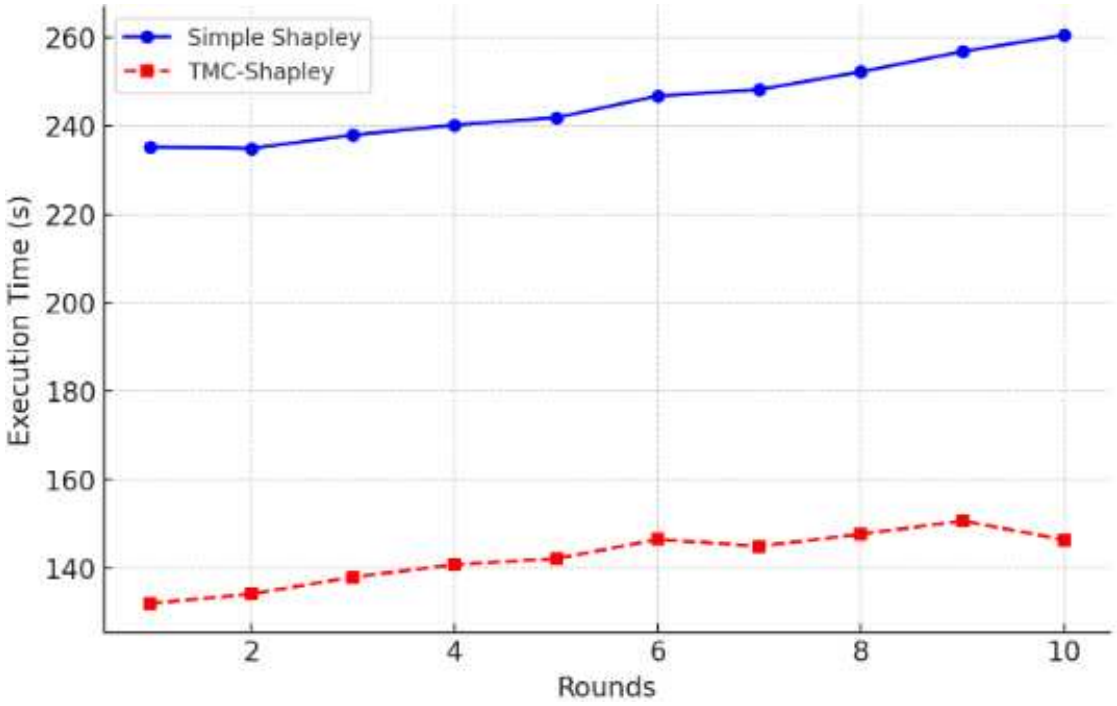

FIGURE 5.6 : Time Comparison : Simple Shapley vs TMC-Shapley (Adult Dataset).

It converges faster, reaching over 85.5% accuracy by round 4, while SV only reaches this point around round 7. Fig. 5.6 highlights the computational benefit. TMC-Shapley consistently requires less time — between 132 and 152 seconds per round — compared to 235 to 260 seconds for SV. The TMC-Shapley method closely approximates the simple SV evaluation in terms of accuracy while reducing the computation time by over 45% on average. Although some individual client contributions differ in magnitude due to random sampling, the overall accuracy improvement is comparable, validating the efficiency of TMC-Shapley.

5.4.2 IMPACT OF REPUTATION FILTERING

Our empirical evaluation reveals important findings on the performance and robustness of our approach under adversarial conditions. The label poisoning attack significantly reduced accuracy after the 70th round. Reputation filtering improved model accuracy linearly, excluding unreliable or adversarial updates. A crucial aspect of our reputation-based filtering mechanism is the selection of the reputation threshold, τ. This threshold directly impacts model security and accuracy, as it determines which client updates are considered trustworthy enough to be aggregated. If τ is set too low, adversarial or low-quality client updates may persist, undermining global model performance. Conversely, an overly strict threshold may exclude legitimate contributors, reducing data diversity and collaborative efficacy.

To select τ, we performed a grid search over $[0.7, 0.95]$ for all benchmark datasets and data partitioning scenarios (S1, S2, S3). For each value, we evaluated (1) the rate of adversarial client exclusion, (2) retention of benign updates, and (3) overall accuracy. For MNIST and Adult, $\tau = 0.85$ yielded an optimal trade-off : adversarial clients were reliably filtered following the attack, while honest clients largely remained active, preserving global

accuracy. For FetalHealth, the optimal τ was slightly lower ($\tau = 0.80$), reflecting higher heterogeneity and the need to preserve data variety.

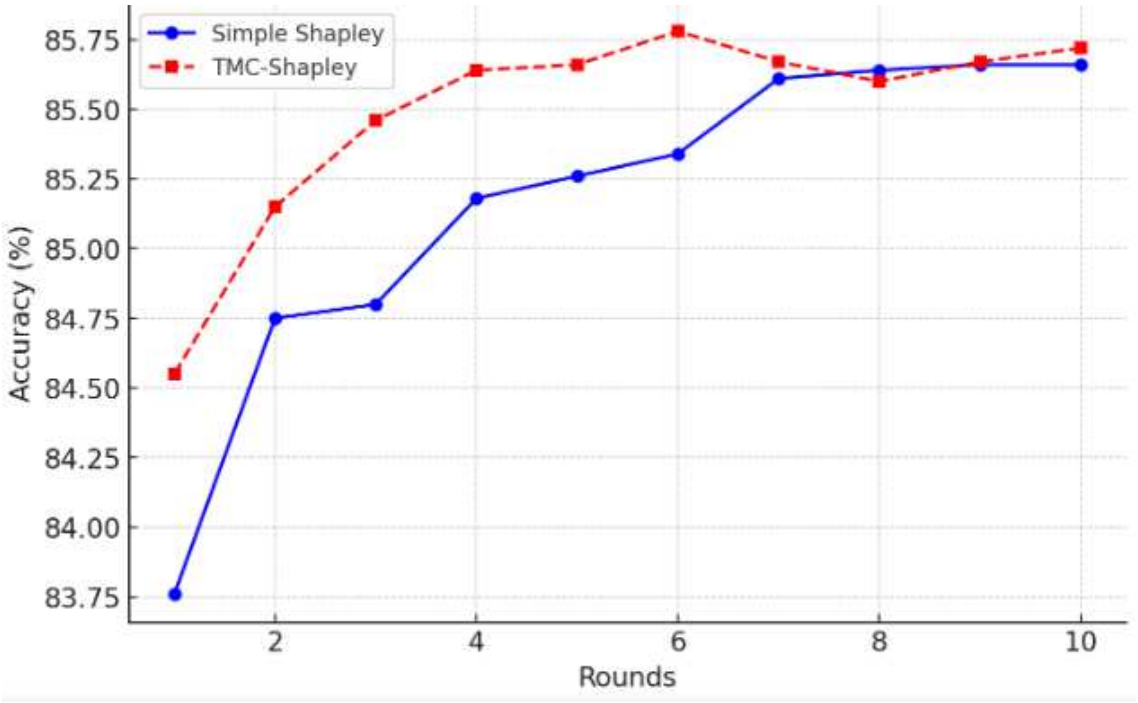

FIGURE 5.7 : Accuracy Comparison : Simple Shapley vs TMC-Shapley (Adult Dataset).

5.4.3 DATASET-DEPENDENT PERFORMANCE

The overall performance of our approach is intrinsically tied to the characteristics of the underlying datasets. MNIST, with its homogeneous and high-quality data, consistently achieves high accuracy across methods. In contrast, the Adult and Fetal Health datasets, which exhibit significant heterogeneity, present a more challenging environment, highlighting the need for robust contribution evaluation and filtering mechanisms. Hence, our approach yields better performance than other proposed approaches.

5.4.4 LARGE SCALE SCALABILITY

Scalability is essential for FL frameworks that target hundreds or thousands of clients. The reputation mechanism used in RBFL is computationally efficient : gradient cosine similarities and reputation score updates only scale linearly with the number of clients ($O(n)$ per round), and operations can be parallelized as they are independent across clients.

The principal computational challenge is data valuation, specifically the estimation of Shapley values. The classic SV has exponential ($O(n!)$) complexity, making it impractical for large n. In contrast, TMC-Shapley enables an efficient approximation ; the number of Monte Carlo permutations, ($\bar{M}$), is tunable, allowing for a balance between computational resources and valuation fidelity. Moreover, for massive-scale real-world deployment, RBFL can be enhanced with a clustering process where similar clients are clustered and perform valuation at the cluster level.

5.4.5 DYNAMIC LEARNING RATE ADAPTATION :

An important design feature of the RBFL is its ability to dynamically adapt the learning rate in each communication round. Specifically, the learning rate is scaled proportionally to the number of participating clients : it increases when fewer clients contribute and decreases when more clients are involved. This ensures that updates remain meaningful and stable across different client activity levels.

In contrast, traditional FL strategies such as FedAvg typically use a fixed learning rate, which can lead to slow learning when client participation is low or instability when too many heterogeneous updates are aggregated. Our adaptive mechanism addresses this by maintaining consistent progress across rounds without manual tuning. This proves especially useful in realistic FL deployments where client availability may fluctuate significantly.

5.5 THREATS TO VALIDITY

In this section, we discuss the possible threats to the validity of the evaluation conducted in this study through the adherence to a specified set of criteria Wohlin *et al.* (2012). We investigated construct, internal reliability, and external validity to guarantee the robustness and generalizability of our findings.

Threats to Construct Validity, concern the correlation between theory and observations. Our study utilizes three benchmark datasets to replicate non-IID and imbalanced data distributions. Although these datasets are used extensively in FL research, they may not adequately represent the richness and diversity of real-world data distributions. The generating procedure of non-IID data may include biases that inadequately represent the diversity of real-world data distributions. A possible concern is the accuracy and reliability of the reputation mechanism. The metrics that assess the data utility may influence the algorithm's efficacy. If the metrics fail to capture the true contribution of the clients accurately, it could impact overall efficiency and accuracy. Our approach may generate errors if it is not well-tuned to address the differing levels of data imbalance across clients.

Threats to Internal Validity, Several factors may affect the internal validity of our findings. The investigation was carried out in a controlled laboratory setting with predetermined client behavior. This may not accurately represent real-world FL scenarios with dynamic client participation and varying data distributions over time. The hyperparameters of our empirical study (e.g., reputation score, learning rate, number of communication rounds, and clients) can substantially influence the outcomes. Small variations could lead to significant differences in detection accuracy. The attack detection and mitigation method is evaluated primarily against label-flipping attacks. Their generalizability to other poisoning attacks (e.g., gradient manipulation or backdoor attacks) remains unverified. Although

RBFL was not directly evaluated under these additional threats in this study, its cosine-similarity-based reputation mechanism naturally penalizes update patterns that diverge from benign behavior, causing adversarial clients' reputation scores to drop and leading to their exclusion. Specifically, gradient manipulation, which is designed to covertly poison the aggregation, is expected to result in low similarity, triggering filtration. For backdoor attacks, additional anomaly detection may be needed in tandem with reputation monitoring.

Reliability Validity Threats, pertains to the feasibility of duplicating this study. To guarantee reproducibility, we provided all essential information on our empirical evaluation, encompassing dataset description, hyperparameter settings, and assessment processes. Furthermore, we will publicly release our source code and the implementation of the adaptive client clustering methodology. This encompasses the FL framework, data partitioning scripts, contribution valuation, reputation mechanism, and security against poisoning attacks employed in our research. All data utilized in this empirical investigation are accessible online [2].

Threats to External Validity, concern the generalizability of our findings to real-world scenarios. While our investigation demonstrates the effectiveness of client clustering in simulated non-IID and imbalanced settings, the results may not generalize to all FL applications. For instance, focusing on a specific set of clients and datasets may not fully represent the heterogeneity of data distributions in practical deployments, such as healthcare or finance. Furthermore, our investigation assumes a controlled environment with semi-reliable client and central server communication. However, clients' behavior in FL systems, such as device heterogeneity, network conditions, and device failures, may vary in real-world applications, potentially limiting the generalizability of our findings to specific contexts. The threat models assume controlled adversarial behavior. In practice,

2. https ://github.com/KingMazu/NC_simple-tmc_shapley

attackers may employ more sophisticated strategies (e.g., adaptive or colluding attacks), which could evade detection.

We plan to conduct further evaluations with more diverse datasets and client configurations to validate the generalizability and robustness of our approach. We will also explore the impact of different network conditions and privacy-preserving techniques on our method's performance.

5.6 CONCLUSION

This chapter (my publication Rehman *et al.* (2025b)) introduced a reputation-based defense mechanism for securing federated learning against data poisoning attacks. This chapter addresses the problem defined in the introduction of the thesis research question 3. Our approach significantly enhances global model accuracy and diminishes computational overhead by integrating a cosine-gradient-based utility measurement with fair contribution assessment through the Shapley Value and its efficient TMC-Shapley approximation. It also incorporates a robust reputation filtering mechanism. Our extensive empirical evaluation on MNIST, Adult, and Fetal Health datasets reveals that our method effectively mitigates the adverse effects of malicious or low-quality client updates, ensuring a fair distribution of contributions and enhanced robustness.

CHAPITRE VI

CONCLUSION

This thesis addresses the issue of safe, secure, and trustworthy AI systems. This thesis integrates our four key research contributions : a survey article regarding safe and intelligent healthcare Rehman *et al.* (2026b), a conceptual framework of safe AI Rehman *et al.* (2026a), along with my two original research studies Rehman *et al.* (2025a,a) that address fairness, robustness, and security in collaborative learning for safety-critical systems. Unitedly, we defined Safe AI and emphasized the critical importance of ensuring its safety in industries such as healthcare, aviation, and finance, where errors might have serious consequences. First, the research commences with a conceptual framework of Safe AI, defined by fairness, robustness, reliability, and alignment with human values. The study examines their significance and related challenges, as well as governmental efforts to ensure AI safety, promoting the development of trustworthy systems.

Secondly, to ensure fairness and robustness, the development of AI models necessitates sufficient training data in terms of quantity and quality to make meaningful predictions. However, training these models requires substantial computational resources and extensive training data. To address the issue of limited training data, researchers utilize collaborative learning, which involves multiple smaller groups pooling their computational resources and local data to train a global model that benefits all contributors. However, it faces two primary challenges : (i) data heterogeneity across subgroups, which undermines fairness and robustness, and (ii) vulnerability to adversarial threats, particularly data poisoning, which compromises security and reliability. In this context, we proposed FEDCIM, an innovative approach to alleviate the impact of heterogeneous data distributions, enhance fairness and robustness through iterative clustering. Although this method shows superior

performance compared to well-known methods like FedAvg and CFL, it incurs higher time and computational costs, which will need to be addressed in future work.

Finally, A reputation-based defence mechanism to protect collaborative learning from data poisoning attacks is also outlined, which improves model accuracy and reduces computational load by integrating utility measurement with fair contribution assessment. Empirical findings indicate its effectiveness in mitigating the negative impacts of malicious client updates, suggesting avenues for future enhancements, including dynamic reputation adjustment and application to more complex adversarial environments.

Collectively, these contributions advance the state of the art in collaborative learning by integrating fairness, robustness, and security into its design. By integrating theoretical insights with practical solutions, this thesis illustrates the realization of Safe AI for safety-critical systems.

CHAPITRE VII
FUTURE WORK

While this thesis makes progress toward improving fairness, robustness, and security in federated learning through the proposed FEDCIM and RBFL framework, several important directions remain open for future exploration.

First, although FEDCIM mitigates the impact of data heterogeneity through iterative clustering, its performance could be further evaluated under more complex and large-scale real-world deployments. Future work may explore adaptive clustering mechanisms that dynamically adjust to evolving data distributions over time, particularly in non-stationary environments where client data changes continuously. Additionally, the time and computational cost for iterative clustering are higher than those of others, which may limit scalability in resource-constrained environments. Future work may explore lightweight clustering strategies, model compression techniques, or asynchronous training schemes to improve efficiency without compromising performance.

Second, the RBFL work primarily focuses on robustness in the presence of heterogeneous data and adversarial threats such as data poisoning. Extending this framework to address a broader range of attacks—including model inversion, backdoor attacks, and inference-time adversaries—would further strengthen the security guarantees of the system. Integrating advanced defence mechanisms, such as anomaly detection or trust-based client weighting, could enhance resilience against sophisticated attackers. Additionally, the dynamic adjustment of reputation thresholds extends our approach to more complex adversarial scenarios and extends applicability to non-gradient-based FL scenarios.

Third, while fairness has been considered at a global level, future research could investigate more granular fairness notions, such as subgroup fairness or personalized fairness constraints for individual clients. This would be particularly valuable in sensitive domains like healthcare, where equitable performance across demographic groups is critical.

Finally, real-world validation in domain-specific applications—such as healthcare, finance, or autonomous systems—would be an essential step toward practical adoption. Collaborations with industry partners and deployment on real federated datasets could provide deeper insights into the usability, reliability, and trustworthiness of the proposed approach in operational settings.

BIBLIOGRAPHIE

Becker, B. & Kohavi, R. (1996). *Adult*. UCI Machine Learning Repository. DOI : https ://doi.org/10.24432/C5XW20.

Belyadi, H. & Haghighat, A. (2021). Chapter 4 - Unsupervised machine learning : clustering algorithms. Dans H. Belyadi & A. Haghighat (Éds.), *Machine Learning Guide for Oil and Gas Using Python* pp. 125–168. Gulf Professional Publishing.

Caton, S. & Haas, C. (2024, 7). Fairness in Machine Learning : A Survey. *ACM Comput. Surv.* doi: 10.1145/3616865

Chen, Y., Li, K., Li, G. & Wang, Y. (2024). Contributions Estimation in Federated Learning : A Comprehensive Experimental Evaluation. Dans *Proceedings of the VLDB Endowment*, Vol. 17, pp. 2077–2090. VLDB Endowment. doi: 10.14778/3659437.3659459

Chen, Y., Yang, X., Qin, X., Yu, H., Chen, B. & Shen, Z. (2020). FOCUS : Dealing with Label Quality Disparity in Federated Learning. Repéré à http://arxiv.org/abs/2001.11359

de campos, D. A., Bernardes, J., Garrido, A., de sá, J. M. & leite and, L. P. (2000). SisPorto 2.0 : A Program for Automated Analysis of Cardiotocograms. *Journal of Maternal-Fetal Medicine*, *9*(5), 311–318.

Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, *29*(6), 141–142. doi: 10.1109/MSP.2012.2211477

Du, R., Xu, S., Zhang, R., Xu, L. & Xia, H. (2023). A dynamic adaptive iterative clustered federated learning scheme. *Knowledge-Based Systems*, *276*, 110741.

Duan, M., Liu, D., Ji, X., Wu, Y., Liang, L., Chen, X., Tan, Y. & Ren, A. (2022, 11). Flexible Clustered Federated Learning for Client-Level Data Distribution Shift . *IEEE Transactions on Parallel & Distributed Systems*, pp. 2661–2674.

Espinoza Castellon, F., Mayoue, A., Sublemontier, J.-H. & Gouy-Pailler, C. (2022). Federated learning with incremental clustering for heterogeneous data. Dans

2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi: 10.1109/IJCNN55064.2022.9892653

Fallah, A., Mokhtari, A. & Ozdaglar, A. (2020). Personalized Federated Learning with Theoretical Guarantees : A Model-Agnostic Meta-Learning Approach. Dans H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Éds.). *Advances in Neural Information Processing Systems*, Vol. 33, pp. 3557–3568. Curran Associates, Inc.

for Economic Cooperation, O. & (OCED), D. (2019). *AI Principal*. Repéré à https://oecd.ai/en/ai-principles

Gabriel, I. (2020). AI Alignment : A Comprehensive Survey. *Minds and Machines*, p. 411–437. doi: https://doi.org/10.1007/s11023-020-09539-2

(GDPR), G. D. P. R. (2018). *Data Protection Principals*. Repéré à https://gdpr-info.eu/chapter-2/

Ghorbani, A. & Zou, J. (2019). What is your data worth ? Equitable Valuation of Data" by Amirata Ghorbani and James Zou. Dans *International Conference on Machine Learning*.

Ghorbani, A. & Zou, J. (2020). Data Shapley : Equitable Valuation of Data for Machine Learning. Dans *International Conference on Machine Learning*, pp. 2242–2251.

Ghosh, A., Chung, J., Yin, D. & Ramchandran, K. (2020). An efficient framework for clustered federated learning. Dans *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ghosh, B., Basu, D., Huazhu, F., Yuan, W., Kanagavelu, R., Peng, J. J., Yong, L., Rick, G. S. M. & Qingsong, W. (2024). Don't Forget What I did ? : Assessing Client Contributions in Federated Learning. Repéré à http://arxiv.org/abs/2403.07151

Goddard, M. (2017). The EU General Data Protection Regulation (GDPR) : European Regulation that has a Global Impact. *International Journal of Market Research*, *59*(6), 703–705. doi: 10.2501/IJMR-2017-050

Gong, B., Xing, T., Liu, Z., Xi, W. & Chen, X. (2024). Adaptive Client Clustering for Efficient Federated Learning Over Non-IID and Imbalanced Data. *IEEE Transactions on Big Data*, *10*(6), 1051–1065.

Government, U. (2024). *1st International Scientific Report on Safety of Advance AI*. Repéré le 2024-05-20, à https://www.gov.uk/government/news/safety-of-advanced-ai-under-the-spotlight-in-first-ever-independent-international-scientific-report

Han, J., Kamber, M. & Pei, J. (2012). 2 - Getting to Know Your Data. Dans J. Han, M. Kamber, & J. Pei (Éds.), *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems pp. 39–82. Boston: Morgan Kaufmann, (third edition).

Holzer, P., Jacob, T. & Kavane, S. (2023). *Dynamically Weighted Federated k-Means*. Repéré à https://arxiv.org/abs/2310.14858

Hsu, H.-Y., Keoy, K. H., Chen, J.-R., Chao, H.-C. & Lai, C.-F. (2023). Personalized Federated Learning Algorithm with Adaptive Clustering for Non-IID IoT Data Incorporating Multi-Task Learning and Neural Network Model Characteristics. *Sensors*, *23*(22).

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C. & Song, D. (2019, 11). Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. VLDB Endow.*, p. 1610–1623.

Kairouz, P., McMahan, H. B. & Avent, B. e. a. (2021). Advances and Open Problems in Federated Learning. *arXiv :1912.04977 [cs, stat]*. arXiv : 1912.04977, Repéré le 2021-09-24, à http://arxiv.org/abs/1912.04977

Kang, J., Xiong, Z., Niyato, D., Xie, S. & Zhang, J. (2019). Incentive Mechanism for Reliable Federated Learning : A Joint Optimization Approach to Combining Reputation and Contract Theory. *IEEE Internet of Things Journal*, pp. 10700–10714.

Kearns, M. & Ron, D. (1999). Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, *11*(6), 1427–1453.

Khraisat, A., Alazab, A., Alazab, M., Jan, T., Singh, S. & Uddin, M. A. (2025, 1).

Securing federated learning : a defense strategy against targeted data poisoning attack. *Discover Internet of Things*, p. 16.

Khuu, D.-P., Sober, M., Kaaser, D., Fischer, M. & Schulte, S. (2024). Data Poisoning Detection in Federated Learning. Dans *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, p. 1549–1558., New York, NY, USA. Association for Computing Machinery.

Kim, Y., Hakim, E. A., Haraldson, J., Eriksson, H., da Silva Jr. au2, J. M. B. & Fischione, C. (2020). *Dynamic Clustering in Federated Learning*. Repéré à https://arxiv.org/abs/2012.03788

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. pp. 32–33.

Kumar, S., Lakshminarayanan, A. & et al. (2022). Towards More Efficient Data Valuation in Healthcare Federated Learning Using Ensembling. p. 119–129., Berlin, Heidelberg. Springer-Verlag. doi: 10.1007/978-3-031-18523-6_12

Lahitani, A. R., Permanasari, A. E. & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure : Study case in online essay assessment. Dans *2016 4th International Conference on Cyber and IT Service Management*, pp. 1–6.

Lai, W., Xu, Z. & Yan, Q. (2024). Clustered Federated Learning Based on Client's Prototypes. Dans *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 909–914.

Li, J., Guo, W., Han, X., Cai, J. & Liu, X. (2022). *Federated Learning based on Defending Against Data Poisoning Attacks in IoT*. Repéré à https://arxiv.org/abs/2209.06397

Li, J., He, J., Hu, L. & Wu, C. (2024, 4). A Federated Learning Framework via Decentralized Data Valuation for Chronic Disease Healthcare. Dans *ACM International Conference Proceeding Series*, pp. 60–65. Association for Computing Machinery. doi: 10.1145/3669828.3669837

Li, T., Hu, S., Beirami, A. & Smith, V. (2021, 18–24 Jul). Ditto : Fair and Robust Federated Learning Through Personalization. Dans M. Meila & T. Zhang (Éds.). *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 de *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR.

Li, T., Hu, S., Beirami, A. & Smith, V. (2021, 18–24 Jul). Ditto : Fair and Robust Federated Learning Through Personalization. Dans M. Meila & T. Zhang (Éds.). *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 de *Proceedings of Machine Learning Research*, pp. 6357–6368. PMLR.

Li, Y., Wang, X. & An, L. (2023). Hierarchical Clustering-based Personalized Federated Learning for Robust and Fair Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *7*.

Liu, Z., Chen, Y., Yu, H., Liu, Y. & Cui, L. (2021). GTG-Shapley : Efficient and Accurate Participant Contribution Evaluation in Federated Learning. Repéré à http://arxiv.org/abs/2109.02053

Liu, Z., Chen, Y., Zhao, Y. & et al. (2022). *Contribution-Aware Federated Learning for Smart Healthcare*. Repéré à https://www.webank.com/

Liu, Z., Guo, J., Yang, W., Fan, J., Lam, K.-Y. & Zhao, J. (2024). Dynamic User Clustering for Efficient and Privacy-Preserving Federated Learning. *IEEE Transactions on Dependable and Secure Computing*, pp. 1–12.

Lu, H., Thelen, A., Fink, O., Hu, C. & Laflamme, S. (2024). Federated learning with uncertainty-based client clustering for fleet-wide fault diagnosis. *Mechanical Systems and Signal Processing*, *210*, 111068.

Luo, J., Mendieta, M., Chen, C. & Wu, S. (2023). PGFed : Personalize Each Client's Global Objective for Federated Learning. Dans *2023 IEEE International Conference on Computer Vision (ICCV)*, pp. 3923–3933.

Lyu, L., Xu, X. & Wang, Q. (2020). Collaborative Fairness in Federated Learning. Repéré à http://arxiv.org/abs/2008.12161

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q. & Yu, P. S. (2022). *Privacy and Robustness in Federated Learning : Attacks and Defenses*. Repéré à https://arxiv.org/abs/2012.06337

McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. (2016). Communication-Efficient Learning of Deep Networks from Decentralized Data. Repéré à http://arxiv.org/abs/1602.05629

Morafah, M., Vahidian, S., Wang, W. & Lin, B. (2022). FLIS : Clustered Federated Learning via Inference Similarity for Non-IID Data Distribution. *arXiv preprint arXiv :2208.09754*.

of Standards, N. I. & (NIST), T. (2021, décembre). *AI Risk Management Framework Concept Paper*. Repéré à https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf

of Standards, N. I. & (NIST), T. (2021, octobre). *Taxonomy of AI Risk*. Repéré à https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf

Ouyang, X., Xie, Z., Zhou, J., Huang, J. & Xing, G. (2021). ClusterFL : a similarity-aware federated learning system for human activity recognition. Dans *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '21, p. 54–66., New York, NY, USA. Association for Computing Machinery.

Park, H.-S. & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2, Part 2), 3336–3341. doi: https://doi.org/10.1016/j.eswa.2008.01.039

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J. & et al (2019). PyTorch : An Imperative Style, High-Performance Deep Learning Library. Dans H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Éds.). *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Qayyum A, Qadir J, B. M. A.-F. A. (2021). Secure and Robust Machine Learning for Healthcare : A Survey. *IEEE Rev Biomed Eng*, *14*, 156–180. doi: https://doi.org/10.1109/RBME.2020.3013489

Radovič, B. & Pejović, V. (2023). *REPA : Client Clustering without Training and Data Labels for Improved Federated Learning in Non-IID Settings*. Repéré à https://arxiv.org/abs/2309.14088

refer, A. C. D. A. *Critical Infrastructure Security and Resilience*. Repéré à https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors

refer, U. H. . H. S. *HIPAA Privacy Rule*. Repéré à https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

Rehman, A., Ameyed, D. & Jaafar, F. (2025, octobre). FedCIM : Handling Data Heterogeneity in Federated Learning to Improve Fairness and Robustness . Dans *2025 IEEE Conference on Dependable, Autonomic and Secure Computing (DASC)*, pp. 98–107., Los Alamitos, CA, USA. IEEE Computer Society. doi: 10.1109/DASC68382.2025.00020

Rehman, A., Ameyed, D. & Jaafar, F. (2026). Chapter 6 : Ensuring Safe and Robust AI : Fairness, Explainability, and System Reliability. Dans *Transparent Intelligent : A guide to Explainable AI*. Nova Science Publishers, Inc.

Rehman, A., Ameyed, D. & Jaafar, F. (2026). Safe and Intelligent Healthcare 5.0 : A Survey to Explore Challenges, Trends, and Innovative Solutions. *ACM Computing Survey*.

Rehman, A., Mazu, I. I., Jaafar, F., Ameyed, D. & Abdessalem, H. B. (2025, octobre). RBFL : Securing Federated Learning against Data Poisoning Using a Reputation-based Approach . Dans *2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–10., Los Alamitos, CA, USA. IEEE Computer Society. doi: 10.1109/AICCSA66935.2025.11315401

Salazar, T., Araujo, H., Cano, A. & Abreu, P. H. (2026). A survey on group fairness in federated learning : challenges, taxonomy of solutions and directions for future research. *Artificial Intelligence Review*, *59*(2), 1573–7462. doi: 10.1007/s10462-025-11475-5

Sattler, F., Müller, K.-R. & Samek, W. (2019). Clustered federated learning. Dans *Proceedings of the NeurIPS'19 Workshop on Federated Learning for Data Privacy and Confidentiality*, pp. 1–5.

Sattler, F., Müller, K.-R. & Samek, W. (2021). Clustered Federated Learning : Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3710–3722.

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, *2*.

Shi, Y., Yu, H. & Leung, C. (2021). Towards Fairness-Aware Federated Learning. doi: 10.1109/TNNLS.2023.3263594

Shu, J., Yang, T., Liao, X., Chen, F., Xiao, Y., Yang, K. & Jia, X. (2023). Clustered Federated Multitask Learning on Non-IID Data With Enhanced Privacy. *IEEE Internet of Things Journal*, *10*, 3453–3467.

Siala, H. & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare : A systematic review. *Social Science Medicine*, *296*, 114782. doi: https://doi.org/10.1016/j.socscimed.2022.114782

Sim, R. H. L., Zhang, Y., Chan, M. C. & Low, B. K. H. (2020). Collaborative Machine Learning with Incentive-Aware Model Rewards. Repéré à http://arxiv.org/abs/2010.12797

Siomos, V. & Passerat-Palmbach, J. (2023). Contribution Evaluation in Federated Learning : Examining Current Approaches. Repéré à http://arxiv.org/abs/2311.09856

Smith, V., Chiang, C.-K., Sanjabi, M. & Talwalkar, A. (2017). Federated multi-task learning. Dans *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4427–4437., Red Hook, NY, USA. Curran Associates Inc.

Sun, H., Tang, X., Yang, C., Yu, Z., Wang, X., Ding, Q., Li, Z. & Yu, H. (2023). Hierarchical Federated Learning Incentivization for Gas Usage Estimation. Repéré à http://arxiv.org/abs/2307.00233

Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J. & Zhang, C. (2022). FedProto : Federated Prototype Learning across Heterogeneous Clients. Dans *AAAI Conference on*

Artificial Intelligence.

Tian, P., Liao, W., Yu, W. & Blasch, E. (2022). WSCC : A Weight-Similarity-Based Client Clustering Approach for Non-IID Federated Learning. *IEEE Internet of Things Journal*, *9*(20), 20243–20256.

Wang, G., Dang, C. X. & Zhou, Z. (2019). Measure Contribution of Participants in Federated Learning. Repéré à http://arxiv.org/abs/1909.08525

Wang, T., Rausch, J., Zhang, C., Jia, R. & Song, D. (2020). A Principled Approach to Data Valuation for Federated Learning. Repéré à http://arxiv.org/abs/2009.06192

Winther, R. & Fredriksen, R. (2025, 01). Safe AI vs Safe Use of AI. Dans *Proceedings of the 35th European Safety and Reliability the 33rd Society for Risk Analysis Europe Conference*, pp. 1364–1371. doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7886-cd

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B. & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Wu, Z., Xu, X., Sim, R. H. L., Shu, Y., Lin, X., Agussurja, L., Dai, Z., Ng, S. K., Foo, C. S., Jaillet, P., Hoang, T. N. & Low, B. K. H. (2024). Data valuation in federated learning (pp. 281–296). Elsevier.

Xiao, H., Rasul, K. & Vollgraf, R. (2017). *Fashion-MNIST : a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Repéré à https://arxiv.org/abs/1708.07747

Xiao, Y., Shu, J., Jia, X. & Huang, H. (2021). Clustered Federated Multi-Task Learning with Non-IID Data. Dans *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 50–57.

Xu, X. & Lyu, L. (2020). A Reputation Mechanism Is All You Need : Collaborative Fairness and Adversarial Robustness in Federated Learning. Repéré à http://arxiv.org/abs/2011.10464

Xu, X., Lyu, L., Ma, X., Miao, C., Foo, C. S., Kian, B. & Low, H. (2021). Gradient-Driven Rewards to Guarantee Fairness in Collaborative Machine Learning. *Advances in Neural Information Processing Systems*, *34*, 16104–16117.

Xu, X., Wu, Z., Foo, C. S. & Low, B. K. H. (2021). Validation Free and Replication Robust Volume-based Data Valuation. Dans M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Éds.). *Advances in Neural Information Processing Systems*, Vol. 34, pp. 10837–10848. Curran Associates, Inc.

Xue, Y., Niu, C., Zheng, Z., Tang, S., Lyu, C., Wu, F. & Chen, G. (2021). *Towards Understanding the Influence of Individual Clients in Federated Learning*. Repéré à https://drive.google.com/file

Yan, Y., Tong, X. & Wang, S. (2024). Clustered Federated Learning in Heterogeneous Environment. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(9), 12796–12809.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2019). *Federated Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers. Repéré à https://books.google.ca/books?id=JdPGDwAAQBAJ

Yin, C. & Zeng, Q. (2024). Defending Against Data Poisoning Attack in Federated Learning With Non-IID Data. *IEEE Transactions on Computational Social Systems*, *11*(2), 2313–2325. doi: 10.1109/TCSS.2023.3296885

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S. & Wu, C. (2022, 17–23 Jul). Federated Learning with Label Distribution Skew via Logits Calibration. Dans K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Éds.). *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162 de *Proceedings of Machine Learning Research*, pp. 26311–26329. PMLR.

Zhang, M., Zhao, H., Ebron, S. & Yang, K. (2024). Towards Fair, Robust and Efficient Client Contribution Evaluation in Federated Learning. Repéré à http://arxiv.org/abs/2402.04409

Zhao, J., Zhu, H., Wang, F., Zheng, Y., Lu, R. & Li, H. (2024). Efficient and Privacy-Preserving Federated Learning Against Poisoning Adversaries. *IEEE Transactions on*

Services Computing, *17*(5), 2320–2333. doi: 10.1109/TSC.2024.3377931

Zhu, S., Zeng, J., Wang, S., Sun, Y., Li, X., Yao, Y. & Peng, Z. (2024). *On ADMM in Heterogeneous Federated Learning : Personalization, Robustness, and Fairness*. Repéré à https://arxiv.org/abs/2407.16397

Zhu, S., Zeng, J., Wang, S., Sun, Y., Li, X., Yao, Y. & Peng, Z. (2024). *On ADMM in Heterogeneous Federated Learning : Personalization, Robustness, and Fairness*. Repéré à https://arxiv.org/abs/2407.16397

Zhu, Z., Hong, J. & Zhou, J. (2021, 18–24 Jul). Data-Free Knowledge Distillation for Heterogeneous Federated Learning. Dans M. Meila & T. Zhang (Éds.). *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 de *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR.